Actes des 27es Journées Francophones d'Ingénierie des Connaissances

IC 2016

Nathalie Pernelle
Présidente du comité de programme

Sandra Bringay
Présidente du comité d'organisation

6-10 juin 2016, Montpellier, France

AfIA
Association française
pour l'Intelligence Artificielle

Comités

Comité d'organisation

Présidente : Sandra Bringay, UPVM, LIRMM UM-CNRS

- Amin Abdaoui, LIRMM UM-CNRS
- Jérôme Azé, LIRMM UM-CNRS
- Erick Cuenca, LIRMM UM-CNRS
- Samiha Fadloun, LIRMM UM-CNRS
- Dino Ienco, Irstea, TETIS
- Vijay Ingalalli, LIRMM UM-CNRS
- Clément Jonquet, LIRMM UM-CNRS
- Lynda Khiali, TETIS & LIRMM UM-CNRS
- Pierre Larmande, IRD
- Jessica Pinaire, LIRMM UM-CNRS
- Pierre Pompidor, LIRMM UM-CNRS
- Pascal Poncelet, LIRMM UM-CNRS
- Mathieu Roche, Cirad, TETIS
- Arnaud Sallaberry, LIRMM UM-CNRS, UPVM
- Mike Donald Tapi Nzali, LIRMM UM-CNRS
- Maguelonne Teisseire, Irstea, TETIS
- Sarah Zenasni, TETIS & LIRMM UM-CNRS

Comité de programme

Présidente : Nathalie Pernelle, LRI, Université Paris-Sud, France

- Marie-Hélène Abel, Université de Technologie de Compiègne - Heudiasyc, France
- Xavier Aimé, INSERM, France
- Yamine Ait Ameur, INPT-ENSEEIHT, Toulouse - IRIT, France
- Nathalie Aussenac-Gilles, CNRS - IRIT, France
- Bruno Bachimont, Université de Technologie de Compiègne - Heudiasyc, France
- Jean-Paul Barthes, Université de Technologie de Compiègne - Heudiasyc, France
- Aurélien Bénel, Université de Technologie de Troyes - ICD/Tech-CICO, France
- Nacera Bennacer, Supélec, France
- Sandra Bringay, UPVM, LIRMM UM-CNRS, France
- Bertrand Braunschweig, INRIA - Rennes Bretagne Atlantique
- Patrice Buche, INRA- IAT/LIRMM, France
- Elena Cabrio, INRIA, Sophia Antipolis, France

- Jean-Pierre Cahier, Université de Technologie de Troyes - ICD/Tech-CICO, France
- Sylvie Calabretto, Institut National des Sciences Appliquées de Lyon - LIRIS, France
- Gaoussou Camara, Université Alioune Diop de Bambey, Sénégal
- Pierre-Antoine Champin, Université Claude Bernard Lyon 1 - LIRIS, France
- Jean Charlet, Université Pierre et Marie Curie - INSERM U1142 - LIMICS, France
- Olivier Corby, INRIA Sophia Antipolis, France
- Amélie Cordier, LIRIS - Université Claude Bernard Lyon 1 - LIRIS, France
- Olivier Curé, IGM, Université de Marne la vallée, France
- Mathieu D'Aquin, The Open University , Grande Bretagne
- Jérôme David, Université Pierre-Mendès-France - LIG, France
- Sylvie Després, Université Paris 13 - LIMICS, France
- Rim Djedidi, Université Paris 13 - LIMICS, France
- Jean-Pierre Evain, EBU, Suisse
- Gilles Falquet, Université de Genève - Laboratoire ISI, Suisse
- Catherine Faron Zucker, Université Nice Sophia Antipolis - I3S, France
- Cécile Favre, ERIC - Université de Lyon 2, France
- Béatrice Fuchs, Université Jean Moulin Lyon 3 - LIRIS, France
- Frédéric Furst, MIS - Université de Picardie Jules Verne, France
- Jean-Gabriel Ganascia, LIP6 - Université Pierre et Marie Curie, France
- Fabien Gandon, INRIA Sophia Antipolis, France
- Catherine Garbay, LIG - CNRS, France
- Faiez Gargouri, Université de Sfax - ISIMS-Miracl, Tunisie
- Serge Garlatti, Telecom Bretagne - Labsticc, France
- Alain Giboin, INRIA Sophia Antipolis, France
- Monique Grandbastien, LORIA - Université de Lorraine, France
- Christophe Gueret, Data Archiving and Networked Services - DANS/KNAW, Pays-Bas
- Ollivier Haemmerlé, Université de Toulouse Jean Jaurès - IRIT, France
- Mounira Harzallah, LINA - Université de Nantes, France
- Nathalie Hernandez, Université de Toulouse Le Mirail - IRIT, France
- Liliana Ibanescu, AgroParistech - INRA, France
- Antoine Isaac, Europeana & Vrije Universiteit Amsterdam, Pays-Bas
- Marie-Christine Jaulent, INSERM U1142 - LIMICS, France
- Clément Jonquet, LIRMM UM-CNRS, France
- Nadjet Kamel, Université de Sétif, Algérie
- Gilles Kassel, Université de Picardie Jules Verne - MIS, France
- Khaled Khelif, Airbus Defence and Space, France
- Pascale Kuntz, Université de Nantes - LINA, France
- Florence Le Ber, Université de Strasbourg /ENGEES - ICube, France
- Michel Leclère, LIRMM UM-CNRS, France
- Alain Léger, Orange Labs - Rennes, France
- Dominique Lenne, Université de Technologie de Compiègne, Heudiasyc, France
- Moussa Lo, Université Gaston Berger, Sénégal
- Nada Matta, Université de Technologie de Troyes - ICD/Tech-CICO, France

- Alain Mille, Université Claude Bernard Lyon 1- LIRIS, France
- Pascal Molli, Université de Nantes - LINA, France
- Alexandre Monnin, INRIA Sophia-Antipolis, France
- Amedeo Napoli, Université de Lorraine - LORIA, France
- Emmanuel Nauer, Université de Lorraine - LORIA, France
- Jérôme Nobecourt, Université Paris 13 - LIMICS, France
- Alexandre Passant, MDG WEB LIMITED, Irlande
- Cédric Pruski, LIST, Luxembourg
- Yannick Prié, Université de Nantes - LINA, France
- Sylvie Ranwez, Ecole des Mines d'Ales - LGI2P, France
- Chantal Reynaud, LRI, Université Paris-Sud, France
- Catherine Roussey, Irstea, France
- Fatiha Sais, LRI, Université Paris-Sud, France
- Pascal Salembier, Université de Technologie de Troyes - ICD/Tech-CICO, France
- Hassina Seridi, Université Badji Mokhtar-Annaba, Algérie
- Karim Sehaba, Université Lumière Lyon 2 - LIRIS, France
- Nathalie Souf, Université Paul Sabatier Toulouse - IRIT, France
- Sylvie Szulman, LIPN - Université Paris 13 Sorbonne Paris Cité
- Andrea Tettamanzi, Université Nice Sophia Antipolis - I3S, France
- Yannick Toussaint, INRIA Nancy Grand-Est - LORIA, France
- Cassia Trojahn, Université de Toulouse Le Mirail - IRIT, France
- Raphael Troncy, EURECOM, France
- Serena Villata, INRIA Sophia Antipolis, France
- Amel Yessad, Université Paris 6 - LIP6, France
- Haifa Zargayouna, Université Paris 13 - LIPN, France
- Antoine Zimmermann, École des Mines de Saint-Étienne, France
- Pierre Zweigenbaum, CNRS - LIMSI, France

Avant-propos

La sélection d'articles publiés dans ce recueil constitue les actes des 27èmes journées francophones d'Ingénierie des Connaissances (IC) qui se sont déroulées du 6 au 10 juin 2016 à Montpellier.

Cette conférence est le rendez-vous privilégié de la communauté francophone qui s'intéresse aux problématiques liées à l'ingénierie des connaissances. Chercheurs académiques, industriels et étudiants s'y retrouvent pour échanger sur des thématiques de recherche propres à l'acquisition, à la représentation ou à la gestion des données et des connaissances. Ces 27èmes journées francophones d'Ingénierie des Connaissances ont été organisées sous l'égide du collège IC de l'AFIA.

À l'heure du numérique, les données et les outils se multiplient. Cependant, assurer un accès intelligent aux données reste un défi et ce malgré les langages et les technologies qui sont maintenant à disposition des informaticiens et des experts de domaine. Partager des données et des connaissances au sein d'une communauté, d'une entreprise ou sur le web suppose leur explicitation, leur représentation, leur mise en relation, leur diffusion, leur maintenance. L'ingénierie des connaissances est au cœur de ces problématiques.

Parmi trente et un articles soumis, d'auteurs provenant de 4 pays différents, quatorze articles longs et sept articles courts ont été retenus (67% d'acceptation). De plus, une sélection de six posters et trois démonstrations sont présentées dans une session spéciale et sont inclus dans les actes. La soumission des articles a été effectuée de façon anonyme pour permettre une évaluation aussi juste que possible. Dans cette édition, les thèmes abordés par les auteurs couvrent différents aspects de l'ingénierie des connaissances. Ainsi, les approches décrites s'intéressent à la représentation des connaissances, à la recherche d'information ou à l'extraction de connaissances à partir de données plus ou moins structurées mais également aux problématiques liées au partage de l'information et à la détection de communautés. Trois personnalités dont les travaux sont liés à l'ingénierie des connaissances ont accepté de présenter certains aspects de leurs domaines de recherche lors de cette conférence. Mathieu d'Aquin (KMI, Open University) a abordé les architectures ouvertes, distribuées et intelligentes de partage d'information. Mathieu Roche (TETIS, CIRAD) a exposé sa vision de la science des données textuelles. Enfin, Sébastien Mustière (LASTIG, IGN) a présenté comment appréhender l'hétérogénéité de représentation des données géographiques.

Je remercie vivement les auteurs pour leurs contributions, les conférenciers invités pour avoir honoré la conférence de leur présence, les organisateurs des ateliers et tutoriels pour leur investissement ainsi que le comité de programme et les relecteurs additionnels pour la qualité de leurs relectures. Je remercie chaleureusement le comité d'organisation pour son travail efficace et tout particulièrement sa présidente, Sandra Bringay avec qui cela ne peut être qu'un grand plaisir de collaborer. J'en profite également pour remercier Jérôme Azé pour son aide précieuse pour construire ces actes et Jérôme Nobecourt pour m'avoir aidé à constituer la session spéciale dédiée aux posters et démonstrations. Enfin, je remercie l'AFIA pour son soutien à la conférence ainsi que le bureau du collège IC de l'AFIA, et tout particulièrement Jean Charlet, pour m'avoir permis de jouer le rôle de présidente du comité de programme d'IC 2016 et pour m'avoir soutenu dans cette mission.

Nathalie Pernelle
Présidente du comité de programme

Sommaire

Web de Données, Intégration de données et Objets connectés

Détection de consensus entre sources et calcul de confiance fondé sur l'intégrale de Choquet

Fabien Amarger[1,2], Jean-Pierre Chanet[1], Romain Guillaume[2], Ollivier Haemmerlé[2], Nathalie Hernandez[2], Catherine Roussey[1]

[1] UR TSCF, Irstea, 9 av. Blaise Pascal CS 20085, 63172 Aubière,France
prénom.nom@irstea.fr

[2] IRIT, UMR 5505, Université de Toulouse, UT2J 5,allées Antonio Machado, F-31058 Toulouse
prénom.nom@univ-tlse2.fr

Résumé : Aujourd'hui de nombreux entrepôts sont disponibles sur le Web de données liées pour un même domaine d'intérêt. Ces entrepôts peuvent être de qualité variable ce qui rend difficile leur réutilisation. Dans cet article, nous présentons une approche permettant d'identifier la connaissance partagée par différents entrepôts en favorisant la connaissance issue de sources de qualité. L'approche repose sur l'utilisation de l'intégrale de Choquet. Notre approche a été évaluée dans le domaine de l'agriculture.

Mots-clés : ontologies, bases de connaissances, consensus, fusion, fonction de confiance, intégrale de Choquet

1 Introduction

Les technologies du Web sémantique sont maintenant suffisamment matures pour permettre la publication de données structurées sur le Web, contribuant ainsi au Web de données liées. Le Web de données liées doit actuellement faire face à un défi de taille car de plus en plus de données y sont publiées sans indication de leur qualité. Il devient donc difficile de réutiliser ces données. De plus, de nombreux jeux de données sont publiés sur un même domaine. Ces jeux de données mis en ligne par des organismes différents ont souvent été constitués pour répondre à un ou des usages spécifiques. La FAO [1] propose par exemple sur le Web de données liées le thésaurus Agrovoc. Ce thésaurus est utilisé pour cataloguer toute ressource documentaire en lien avec l'agriculture. Les instituts de recherche français comme l'INRA [2] ou l'Irstea [3] ont également développé leur propre thésaurus pour cataloguer les articles scientifiques dans le domaine de l'agriculture. Parallèlement, le projet Agronomic Linked Data propose lui aussi plusieurs ontologies pour faciliter l'intégration de données hétérogènes dans le domaine de la biologie des plantes. Exploiter ces jeux de données pour un nouvel usage implique une analyse approfondie des éléments qui les composent ainsi que de déterminer la qualité des données.

Cet article présente une méthode de construction de bases de connaissances (ontologies avec ou sans individus) qui réutilise simultanément plusieurs bases de connaissances sources (BCS) de qualité variable. De cette manière, il devient possible d'exploiter les éléments communs ainsi que la complémentarité des sources tout en tenant compte des spécificités de chacune d'elles. Chaque élément extrait des différentes sources se voit attribuer un score de confiance. Nos travaux reposent sur l'hypothèse suivante :

1. Food and Agriculture Organization of the United Nations
2. Institut National de la Recherche Agronomique
3. Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture

La confiance d'un élément extrait des sources est fonction de deux critères : (1) le nombre de sources dans lesquelles il apparaît et (2) la qualité de ces sources.

L'article est organisé de la façon suivante. Dans un premier temps, nous dressons un panorama des travaux portant sur la fusion de bases de connaissances. Nous présentons ensuite notre approche de fusion puis nous détaillons nos propositions visant à calculer le score de confiance d'un élément. Finalement, nous présentons une évaluation de notre approche dans le domaine de l'agriculture. Dans la suite de cet article, les exemples illustrant l'approche sont issus de trois sources : le thésaurus Agrovoc de la FAO, la taxonomie des organismes vivants du NCBI[4] et la taxonomie française de référence TaxRef du Muséum d'Histoire Naturelle.

2 État de l'art sur la fusion de bases de connaissances

Construire une base de connaissances à partir de plusieurs BCS existantes est équivalent à un processus de fusion. Nous considérons dans cet article la définition de *fusion* telle qu'elle est proposée dans les travaux Pottinger & Bernstein (2003) :

> En considérant deux modèles A et B et un ensemble de correspondances Map_{AB} établies entre ces deux modèles, le processus de fusion génère un troisième modèle représentant l'union sans doublon des modèles de A et B conformément aux correspondances de Map_{AB}.

Cette définition est suffisamment générique pour considérer comme modèle plusieurs types de sources, dont les ontologies ou les bases de connaissances. La notion d' "union sans doublon" est particulièrement intéressante car elle impose de mettre en place un traitement particulier pour les éléments communs aux deux modèles. Les travaux les plus anciens et les plus emblématiques sont ceux du projet Prompt Noy & Musen (2003). Nous sommes intéressés par les travaux de fusion capables de générer automatiquement une nouvelle base de connaissances contenant les parties communes des sources. Pour comparer les travaux traitant de fusion de bases de connaissances, nous avons défini trois critères :

symétrique : la notion de fusion symétrique implique que les deux modèles à fusionner ont la même importance. Il est aussi possible d'utiliser une technique de fusion asymétrique pour privilégier un modèle plutôt qu'un autre. Dans ce cas, le résultat de la fusion suivra l'organisation du modèle privilégié ; dans notre cas, les sources ont été restructurées en fonction d'un modèle commun Amarger *et al.* (2015).

align : un processus d'alignement génère les correspondances utilisées dans la fusion des modèles. Certains travaux incluent le calcul de l'alignement dans la fusion (inclus) alors que d'autres considèrent l'alignement comme une entrée du processus (entrée) ; il existe de nombreux systèmes d'alignement. Il n'est ici pas nécessaire de proposer un nouveau système mais plutôt de réutiliser des travaux existants et éprouvés. Par conséquent, nous sommes intéressés par les travaux qui dissocient la fusion du calcul d'alignements.

confiance : suivant le processus de fusion appliqué, une confiance peut être associée aux éléments du modèle résultat de la fusion. Nous sommes intéressés par les systèmes de fusion capables de calculer des degrés de consensus entre sources.

4. National Center for Biotechnology Information

Approche	Symétrique	Align.	Confiance
Curé (2009)	sym	entrée	non
Guzmán-Arenas & Cuevas (2010)	sym	inclus	non
Raunich & Rahm (2014)	asym	entrée	non

TABLE 1 – Travaux sur la fusion

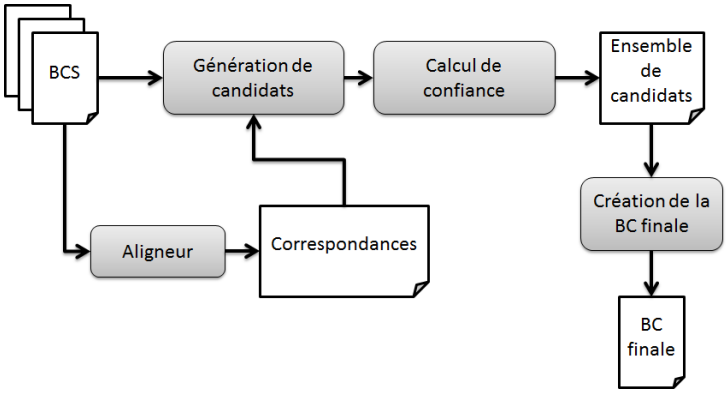


FIGURE 1 – Processus de fusion des bases de connaissances

Nous pouvons remarquer sur le tableau 1 que seule l'approche présentée dans (Raunich & Rahm, 2014) propose un processus asymétrique de fusion. Ce processus asymétrique définissant un modèle prioritaire par rapport à l'autre, certaines ambiguïtés susceptibles d'apparaître lors de la fusion peuvent être résolues automatiquement. Nous remarquons surtout que les processus de fusion présentés ici ne traitent pas de la notion de confiance. On note que les travaux de Guzmán-Arenas & Cuevas (2010) proposent une fonction de confusion pour évaluer la disparité entre deux contraintes de domaine incompatibles. La problématique de fusion étant toujours considérée entre deux modèles, la prise en compte de la confiance à accorder à un élément est simplifiée puisqu'il n'y a que deux possibilités et que la notion d'asymétrie permet de faire un choix dans ce cas-là. Néanmoins, puisque nous considérons que la fusion porte sur plus de deux sources (ce qui est plus réaliste à l'échelle du Web de données liées), nous souhaitons généraliser la notion d'asymétrie en considérant l'importance relative de chaque source dans ce processus. Nous proposons plus précisément de quantifier la confiance à accorder à un élément extrait d'une source en fonction du nombre de sources dans lesquelles il est présent, ainsi que la qualité de chacune d'elles.

3 Processus de fusion de bases de connaissances

Le processus de fusion que nous proposons est décomposé en plusieurs étapes présentées dans la figure 1. Ce processus prend en entrée les différentes BCS à fusionner et fournit en sortie une liste d'éléments pondérés, intitulés candidats, éléments potentiels de la base de connaissances finale. Ces candidats sont sélectionnés à partir d'un seuil pour être intégrés à la base de connaissance finale.

Quatre étapes sont présentes dans ce processus :

alignement des bases de connaissances sources : cette étape établit des alignements entre tous les couples de BCS considérés ;

génération de candidats : à partir des alignements, des candidats sont générés ;

calcul de la confiance : un score de confiance est calculé et associé à chaque candidat ;

construction de la BC : une sélection des candidats est effectuée à partir de leur score de confiance pour déterminer ceux qui appartiendront à la base de connaissance finale. Un filtre automatique peut être complété par une validation manuelle. Ensuite, une fois les candidats sélectionnés, il faut construire pour chacun d'eux l'élément le représentant dans la BC finale (choix de l'URI, choix des métadonnées, choix des labels, etc.).

3.1 Génération des correspondances

Nous définissons une base de connaissances source comme un graphe S composé d'un ensemble de sommets et un ensemble d'arcs $S = (V_S, E_S)$ tels que :

— V_S est l'ensemble des sommets de S. Les sommets sont les classes, les individus et les littéraux de la BCS ;

— E_S est l'ensemble des arcs de S. Les arcs sont toutes les propriétés utilisées pour lier les individus, les classes et les littéraux.

La première étape du processus de fusion est l'alignement entre les différentes BCS. Conformément au fonctionnement des aligneurs, nous effectuons cet alignement entre chaque paire de BCS. Pour chaque paire, nous obtenons un alignement qui est un ensemble de correspondances. Dans cet article, nous ne considérons comme correspondances que les relations d'équivalence stricte ($\equiv$) entre deux sommets appartenant respectivement à chacune des deux sources alignées, c'est-à-dire deux BCS $Si = (V_{Si}, E_{Si})$ et $Sj = (V_{Sj}, E_{Sj})$. Cette relation est pondérée par un degré de fiabilité (fourni par l'aligneur) représentant la probabilité que cette équivalence soit correcte.

Une correspondance se définit comme une arête entre deux sommets $\{oe_i \in V_{Si}; oe_j \in V_{Sj}\}$ pondérée par $valueE(oe_i, oe_j)$. Elle remplit les contraintes suivantes :

— $V_{Si} \neq V_{Sj}$ car une correspondance est toujours établie entre deux sommets appartenant à des ensembles de sommets de BCS différentes (Si et Sj) ;

— une correspondance est toujours établie entre deux sommets de même nature (soit des individus, soit des classes) ;

— $valueE()$ est une application qui, à toute arête définie comme correspondance, associe un unique degré de fiabilité compris entre 0 et 1 tel que $valueE(oe_i, oe_j) = valueE(oe_j, oe_i)$.

Dans nos travaux, nous utilisons l'aligneur LogMap [5] car ce système a obtenu de bons résultats lors de l'évaluation OAEI 2014 (Dragisic *et al.*, 2014). De plus, cet aligneur permet de mettre en correspondance des individus et pas seulement des classes (Jiménez-Ruiz & Grau, 2011). Il n'existe pas à l'heure actuelle d'aligneur capable de générer des correspondances entre propriétés. En conséquence, dans la suite de cet article, par soucis de simplification, nous ne travaillerons que sur la fusion des sommets des graphes représentant les BCS.

5. `http://www.cs.ox.ac.uk/isg/projects/LogMap/`

3.2 Candidat

Les candidats sont générés en exploitant les correspondances établies entre les sommets des graphes Si représentant les différentes BCS.

Un candidat $C = (V_C, E_C, valueE_C)$ est un graphe non-orienté connexe dont les sommets sont des sommets provenant de BCS différentes et les arêtes sont les correspondances issues des T alignements entre les N BCS. Un candidat est un sous-graphe du multigraphe construit à partir des N bases de connaissances sources alignées. Les composants d'un candidat respectent les contraintes suivantes :

— V_C : $\forall v \in V_C$ avec $v \in V_{Si}$ $\nexists v' \in V_C$ tel que $v' \in V_{Si}$ et $v \neq v'$. Tous les sommets d'un candidat appartiennent à des BCS différentes. Par conséquent $|V_C| \leq N$;

— E_C : l'ensemble des arêtes d'un candidat est inclus dans l'ensemble des arêtes des T alignements. Les arêtes de C sont des correspondances ;

— un candidat est un graphe connexe. $\forall v_1, v_2 \in V_C$, il existe forcément un chemin $path = \{e_j, ..., e_k\}$ avec $\forall e_i, e_i \in path, e_i \in E_C$ reliant v_1 à v_2. Par conséquent, tous les sommets de C sont liés à au moins un autre sommet de C par une correspondance.

La figure 2 présente deux candidats liant des individus issus de 3 BCS. Les deux candidats représentent donc des éléments potentiels de la base de connaissances finale, ici "Triticum" et "Triticum Durum".

La génération des candidats équivaut à chercher les composantes connexes dans le graphe global constitué des N BCS alignées. Nous recherchons les composantes de taille inférieure ou égale à N. Nous vérifions que chaque sommet de la composante appartienne à des BCS différentes. Nous effectuons un parcours en profondeur du graphe global en testant les contraintes précédentes. Nous étiquetons les sommets avec l'identifiant du candidat pour éviter les boucles infinies Amarger (2015).

4 Calcul de confiance d'un candidat

Une fois les candidats générés, nous leur affectons un score de confiance. Le premier critère de notre hypothèse cherche à favoriser les candidats contenant le plus grand nombre de sommets issus des différentes BCS et identifiés par l'aligneur comme étant équivalents. Nous avons défini dans Amarger *et al.* (2014) une première fonction $trust_{simple}$ qui prend en compte le nombre de sources impliquées dans le candidat. Cette fonction n'intègre pas le degré de fiabilité accordé par l'aligneur pour la correspondance. Nous définissons dans cet article une autre fonction intitulée $trust_{degree}$ décrite ci-dessous.

Le deuxième critère de notre hypothèse consiste à tenir compte de la qualité des BCS dans le calcul de la confiance d'un candidat. Nous proposons, par la fonction $trust_{choquet}$, de prendre en compte l'implication relative de chaque source pour un candidat donné.

4.1 Fonction Trust Degree

Les candidats sont générés à partir de plus ou moins de correspondances. La fonction $trust_{degree}$ évalue la confiance proportionnellement au nombre de correspondances et à leur degré de fiabilité. Le calcul de cette fonction est présenté dans l'équation 1. $C = (V_C, E_C, valueE)$

est le candidat étudié. E_C est l'ensemble des arêtes du candidat, N le nombre de sources alignées et $valueE$ l'application qui, à chaque arête de E_C, associe son degré de fiabilité.

$$trust_{degree}(C) = \frac{\sum_{e_i \in E_C} valueE(e_i)}{\frac{N(N-1)}{2}} \quad (1)$$

Cette fonction fait la somme de tous les degrés de fiabilité des correspondances utilisées pour générer le candidat. Cette somme est normalisée en divisant le résultat par le nombre maximum de correspondances possibles, c'est-à-dire le nombre de paires possibles entre toutes les sources considérées. Les correspondances étant utilisées dans le processus de génération des candidats, cette fonction permet de prendre en compte indirectement le nombre d'éléments présents dans le candidat. Cette fonction est proportionnelle au nombre d'arêtes : plus le graphe du candidat contiendra d'arêtes, plus il contiendra de sommets.

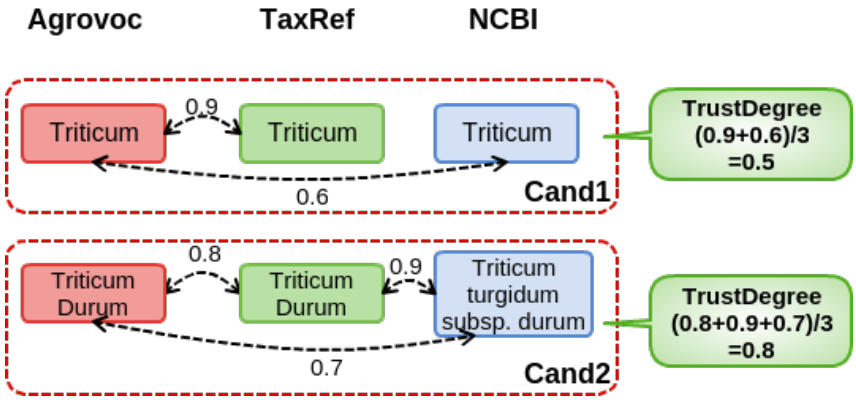


FIGURE 2 – Calcul des scores de confiance avec trustDegree

La figure 2 présente les confiances associées aux deux candidats à l'aide de la fonction $trust_{degree}$. Nous observons que $trust_{degree}(Cand2) > trust_{degree}(Cand1)$. Nous pouvons donc ordonner les candidats par leur confiance, évaluée à l'aide de la fonction $trust_{degree}$ [6].

4.2 Fonction Trust Choquet

L'intégrale de Choquet (eq.2) est utilisée pour la prise de décision sur un ensemble de critères (Grabisch & Roubens (2000)). Elle permet de pondérer l'intérêt de sous-ensembles de critères au lieu de pondérer chaque critère indépendamment des autres, comme le ferait une somme pondérée des critères. Elle permet ainsi de modéliser des conjonctions et des disjonctions sur des sous-ensembles de critères. Pour illustrer l'intérêt de l'intégrale de Choquet, nous allons présenter deux exemples de prises de décision impliquant deux critères x_1 et x_2. La fonction $\mu\{x\}$ représente l'intérêt du critère x dans la prise de décision. Considérons deux cas :

— le cas de la conjonction de critères : le décideur n'est satisfait que si les deux critères x_1 et x_2 sont réalisés simultanément et il n'est pas satisfait si l'un des critères est réalisé sans l'autre. Dans ce cas $\mu(\{x_1\}) = \mu(\{x_2\}) = 0$ mais $\mu(\{x_1, x_2\}) = 1$;

6. Il est à noter que si la confiance était évaluée uniquement par le nombre de sources impliquées, les deux candidats obtiendraient le même score

— le cas de la disjonction : le décideur est satisfait si l'un des deux critères est réalisé sans l'autre. Il n'est pas plus satisfait si les deux critères sont réalisés simultanément. Dans ce cas, $\mu(\{x_1\}) = \mu(\{x_2\}) = 1$ et $\mu(\{x_1, x_2\}) = 1$.

Dans notre cas, les sources S_i impliquées dans le candidat C sont considérées comme les critères de la prise de décision. Une source a un intérêt variable qui dépend du nombre de sources avec lesquelles elle est en accord et de la qualité de ces sources. Par exemple, considérer une nouvelle source pour un candidat impliquant déjà un grand nombre de sources de bonne qualité aura moins d'importance que considérer cette source pour un candidat impliquant des sources de mauvaise qualité. Dans notre cas, la fonction $\mu()$ prendra ses valeurs dans l'intervalle $[0..1]$.

La fonction $f(S)$ retourne une évaluation de la source S.

$$trust_{choquet}(C) = \sum_{i=1}^{n} [f(S_{(i)}) - f(S_{(i-1)})]\mu(A_i) \tag{2}$$

avec $A_i = \{S_{(i)}, ..., S_{(n)}\}$ et $S_{(i)}$ est la permutation des sources S_i tel que $f(S_{(0)}) = 0$ et $0 \leq f(S_{(1)}) \leq f(S_{(2)}) \leq ... \leq f(S_{(n)})$. Cette intégrale calcule le score de confiance du candidat C à partir des sous-ensembles de sources S_i impliquées dans le candidat.

4.2.1 Implication des sources dans un candidat

La fonction f détermine la force de l'implication de la source S dans la construction du candidat C. Cette force est fonction des correspondances associées au sommet de la source S et impliquées dans le candidat C. En effet, nous considérons que plus un sommet de la source S est lié avec des correspondances fortes vers les autres sommets du candidat C, plus la source est impliquée dans la construction de ce candidat.

Prenons l'exemple présenté sur la figure 2 à la page 6. Nous pouvons remarquer pour le candidat "Cand1" que le sommet "Triticum" provenant de la source Agrovoc a une implication plus forte que les autres sommets. En effet, deux correspondances lient le sommet provenant d'Agrovoc vers les deux autres sommets du candidat. Les sommets de sources TaxRef et NCBI n'étant liés qu'à un seul sommet, l'implication de leur source dans le candidat "Cand1" est moindre. Pour représenter formellement cette implication, nous sommons les degrés de fiabilité des correspondances liant le sommet provenant de la source et appartenant au candidat. Afin de normaliser cette valeur, nous divisons cette somme par l'implication maximale possible, c'est-à-dire les correspondances avec un degré de fiabilité de 1 vers tous les sommets du candidat. En d'autres termes, nous la divisons par le nombre de sources considérées moins un. Nous calculons l'implication d'une source S_i par rapport à un candidat sommet C en utilisant l'équation 3. Rappelons qu'une source S est un multigraphe orienté étiqueté S ayant comme ensemble de sommet V_S. Un candidat sommet C est un graphe non-orienté $C = (V_C, E_C, valueE)$.

La fonction f utilisée pour évaluer la source S lors du calcul du score de confiance du candidat C avec l'intégrale de Choquet est définie par la fonction $implication$ comme présenté dans l'équation suivante :

$$f(S) = implication(S, C) = \frac{\sum_{\substack{e \in E_C \\ e=(oe_i, oe_j) \text{ avec } oe_i \in V_S}} valueE(e)}{N - 1} \tag{3}$$

Si nous prenons l'exemple du candidat "Cand1" de la figure 2, l'implication de la source "Agrovoc" dans ce candidat peut être définie de la manière suivante :

$$f(Agrovoc) = implication(Agrovoc, Cand1) = \frac{0,9 + 0,6}{3-1} = \frac{1,5}{2} = 0,75 \quad (4)$$

Alors que l'implication de la source TaxRef dans ce même candidat peut être évaluée de la manière suivante :

$$f(TaxRef) = implication(TaxRef, Cand1) = \frac{0,9}{3-1} = \frac{0,9}{2} = 0,45 \quad (5)$$

Nous ne considérons ici que la correspondance qui a un degré de fiabilité de 0,9 puisque c'est la seule qui implique le sommet provenant de la source TaxRef. De la même manière, nous définissons l'implication de la source NCBI dans le candidat "Cand1" de la manière suivante :

$$f(NCBI) = implication(NCBI, Cand1) = \frac{0,6}{3-1} = \frac{0,6}{2} = 0,3 \quad (6)$$

Cette notion d'implication est particulièrement pertinente dans cet exemple puisque nous observons que le sommet provenant d'Agrovoc est central dans la construction de ce candidat. Les deux autres sommets n'ont pas de correspondance entre eux. Si ce sommet n'était pas présent, alors le candidat n'existerait tout simplement pas. Il est donc cohérent que l'implication de la source Agrovoc soit bien plus grande que l'implication des deux autres sources.

Si un candidat C n'a qu'un seul sommet, et donc aucune correspondance à utiliser, alors nous définissons l'implication de la source S de la manière suivante : $f(S) = implication(S, C) = \frac{1}{N-1}$. Rappelons que N est le nombre de sources alignées dans le processus de fusion.

4.2.2 Intérêt des sources

La deuxième fonction à définir pour utiliser l'intégrale de Choquet est la fonction μ qui représente l'intérêt des sources dans la prise de décision. Cela permet de définir des priorités entre les sources. Nous pouvons par exemple favoriser les candidats impliquant la source "TaxRef" plutôt que ceux impliquant "Agrovoc". Pour ce faire, nous définissons une fonction $Q(S)$ retournant une valeur, comprise entre 0 et 1, représentant la qualité de la source S. L'intérêt d'une source sera fonction de sa qualité.

Dans notre exemple, nous considérons trois sources de qualité différente. Pour évaluer $Q(S)$, nous utilisons les scores de qualité définis avec nos experts lors de la construction de notre référence sur la taxonomie des blés (voir section Expérimentation). La source TaxRef, qui est une référence nationale dans ce domaine, a un score de qualité fixé à 0,9. Du fait de son processus de validation manuel et de sa mise à jour régulière, la source NCBI a également un score relativement élevé fixé à 0,8. La source Agrovoc, quant à elle, a un score de qualité de 0,6 puisque des travaux Soergel *et al.* (2004) ont montré qu'elle contient un certain nombre d'erreurs. Elle reste néanmoins une source intéressante.

Nous devons définir la fonction $\mu(L_S)$ caractérisant l'intérêt d'un sous-ensemble de sources ($L_S = \{S_j, ..., S_k\}$) en fonction de leur qualité. De cette façon, nous pouvons prendre en compte la diversité et la multiplicité des sources. Un candidat impliquant un grand nombre de sources

de mauvaise qualité pourra être considéré aussi pertinent qu'un candidat impliquant peu de sources de très bonne qualité. Nous considérons non seulement que chaque source a un intérêt variable mais aussi que l'évolution de l'intérêt des sources n'est pas linéaire. Cette non-linéarité permet de prendre en compte une évolution variable de l'intérêt des sources. Nous pouvons, par exemple, considérer que si un candidat implique déjà un grand nombre de sources de bonne qualité, alors l'ajout d'une nouvelle source de bonne qualité dans la définition du candidat ne va pas augmenter significativement sa confiance. Notre intuition sur cette répartition non-linéaire est qu'il existe un point représentatif à partir duquel l'intérêt des sources va croître significativement. Ce point d'explosion [7] est spécifique au problème étudié. Il dépend non seulement du nombre de sources considérées mais aussi de la nécessité de favoriser la qualité ou non des sources. De plus, l'intensité de l'explosion est aussi spécifique au problème étudié.

Nous définissons la fonction $\mu(L_S)$ suivante, en considérant L_S comme étant un sous-ensemble des sources considérées :

$$\mu(L_S) = \frac{\lambda(\sum_{i=1}^{|L_S|} Q(S_i)) - \lambda(0)}{\lambda(\sum_{i=1}^{N} Q(S_i)) - \lambda(0)} \tag{7}$$

$$\lambda(x) = arctan(\frac{x - x_0}{\gamma}) \tag{8}$$

La fonction $Q(S_i)$ permet de récupérer la qualité de la source S_i définie précédemment. L'équation $\sum_{i=1}^{N} Q(S_i)$ permet d'obtenir la somme des scores de qualité de toutes les sources considérées dans le processus. Dans l'exemple, nous pouvons définir :

$$\sum_{i=1}^{N} Q(S_i) = 0,9 + 0,8 + 0,6 = 2,3 \tag{9}$$

De la même façon, nous utilisons l'équation $\sum_{i=1}^{|L_S|} Q(S_i)$ qui permet d'obtenir la somme des scores de qualité des sources présentes dans le sous-ensemble L_S.

Cette fonction $\mu(L_S)$ permet de représenter l'intérêt du sous-ensemble de sources L_S. Nous utilisons la fonction $\lambda(x)$ qui est inspirée de la fonction de répartition de la loi gamma et qui permet d'avoir une répartition qui respecte notre intuition sur l'évolution de l'intérêt des sources.

Dans la fonction $\lambda(x)$, deux paramètres sont utilisés. Le premier, x_0, permet de définir le point d'explosion, point à partir duquel l'intérêt des sources augmente particulièrement. Le deuxième paramètre est la valeur γ qui définit l'indice de linéarité de la courbe. Plus la valeur de γ tend vers 0 et plus l'intérêt des sources est crénelé, c'est-à-dire qu'il est très proche de 0 en dessous de x_0 et très proche de 1 au dessus. À l'inverse, plus γ s'approche de $\sum_{i=1}^{N} Q(S_i)$ (somme des scores de qualité de toutes les sources considérées) et plus la courbe est linéaire.

Pour notre cas d'étude, nous devons définir les deux paramètres x_0 et γ. Nous définissons arbitrairement le point d'explosion à 50% de la qualité disponible. Soit $x_0 = 2,3/2 = 1,15$.

Toujours arbitrairement, nous définissons un taux de linéarité de la répartition à 20%. Nous définissons $\gamma = 2,3 * 0,20 = 0,46$.

La figure 3 présente la répartition de la fonction $\mu(x)$ en fonction de nos paramètres.

Comme vu précédemment, si nous considérons le candidat "Cand1" de la figure 2, nous avons les implications des sources suivantes :

7. Point auquel la dérivée μ' est à son maximum

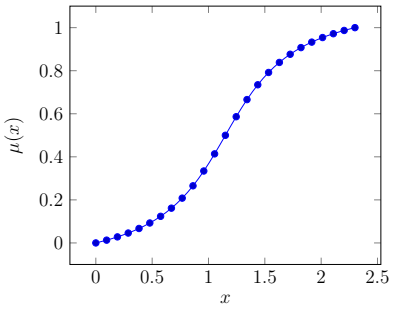


FIGURE 3 – Répartition de $\mu(x)$ avec les paramètres $\sum_{i=1}^{N} Q(S_i) = 2,3$, $x_0 = 1,15$ et $\gamma = 0,46$

— $implication(Agrovoc, Cand1) = 0,75$
— $implication(TaxRef, Cand1) = 0,45$
— $implication(NCBI, Cand1) = 0,3$

Le calcul de la confiance du candidat "Cand1" en utilisant l'intégrale de Choquet est :

$$\begin{aligned} trust_{choquet}(Cand1) =& f(NCBI) * \mu(Agrovoc, TaxRef, NCBI) \\ &+ [f(TaxRef) - f(NCBI)] * \mu(Agrovoc, TaxRef) \\ &+ [f(Agrovoc) - f(TaxRef)] * \mu(Agrovoc) \\ =& 0,3 * \mu(Agrovoc, TaxRef, NCBI) \\ &+ (0,45 - 0,3) * \mu(Agrovoc, TaxRef) \\ &+ (0,75 - 0,45) * \mu(Agrovoc) \\ =& 0,3 * 1 + 0,15 * 0,77 + 0,3 * 0,14 = 0,46 \end{aligned} \tag{10}$$

5 Évaluation

Nous avons construit semi-automatiquement trois bases de connaissances en transformant les 3 sources suivantes : le thésaurus Agrovoc de la FAO, la taxonomie des organismes vivants du NCBI et la taxonomie française de référence TaxRef des organismes vivants du Muséum d'Histoire Naturelle. Le processus de transformation est présenté dans Amarger *et al.* (2014). Nous avons aussi construit une référence sur la taxonomie des blés avec l'aide de trois experts Amarger (2015). Cette référence contient l'union de tous les éléments des 3 BCS précédentes validés manuellement par nos experts.

La précision, le rappel et la $F - mesure$ sont calculés afin d'évaluer la qualité des candidats générés par notre processus de fusion. Nous considérons qu'un candidat généré est valide si la totalité de ses sommets appartiennent à la référence. Nous avons adapté ces mesures de la manière suivante :

— la précision est le rapport entre le nombre de sommets des candidats valides et le nombre total de sommets des candidats générés par la fusion ;
— le rappel est le rapport entre le nombre de sommets des candidats valides et le nombre total de sommets de la référence ;
— la $F - mesure$ est une combinaison des mesures de précision et rappel.

	$trust_{degree}$			$trust_{choquet}$		
Seuil	**Précision**	**Rappel**	**F-Mesure**	**Précision**	**Rappel**	**F-Mesure**
0,1	0,99	0,6	0,77	0,87	0,98	0,92
0,2	0,99	0,63	0,77	0,99	0,63	0,77
0,3	**1**	0,62	0,77	0,99	0,63	0,77
0,4	1	0,60	0,75	1	0,62	0,77
0,5	1	0,60	0,75	1	0,60	0,75
0,6	1	0,60	0,75	1	0,60	0,75
0,7	1	0,26	0,41	1	0,26	0,41
0,8	1	0,26	0,41	1	0,26	0,41
0,9	1	0,26	0,41	1	0,26	0,41

TABLE 2 – Expérimentations $trust_{degree}$ vs $trust_{choquet}$

Nous avons analysé l'impact de chacune des fonctions de confiance en sélectionnant les candidats en fonction de leur score de confiance. Un candidat est sélectionné si son score est supérieur à un seuil. Nous sélectionnons les candidats par pas de 0,1. Nous calculons les 3 mesures sur l'ensemble de candidats générés pour chacun des pas. La fonction de confiance représentant au mieux le consensus aura une mesure de précision d'autant plus élevée que le filtre sera élevé lui aussi. Nous ne présentons ici que les résultats concernant les candidats ne contenant que des sommets de nature individus.

En observant les résultats de la fonction $trust_{degree}$, présentés dans le tableau 2, nous observons plusieurs phénomènes. Tout d'abord, la précision est importante dès le seuil 0,1. Le rappel est moins satisfaisant. Ceci s'explique par le fait que la fonction de confiance $trust_{degree}$ discrimine l'ensemble des candidats en fonction des correspondances. Les candidats ayant un haut score de confiance sont ceux qui impliquent beaucoup de correspondances. Les candidats impliquant moins de 3 sources sont ici rejetés dès les seuils assez bas, bien que certains soient valides. Ceci explique la diminution rapide du rappel. Les candidats impliquant trois sources n'utilisent pas forcément beaucoup de correspondances ou des correspondances avec de faibles degrés de fiabilité. Les sommets des candidats ne sont alors pas fortement connectés, ce qui explique les faibles valeurs du rappel pour des seuils hauts.

Nous pouvons en déduire que l'utilisation des correspondances dans le calcul de la confiance discrimine rapidement les candidats. Les candidats n'impliquant pas beaucoup de sources auront un score de confiance assez bas. Néanmoins, nous vérifions la validité de notre hypothèse initiale. Plus un candidat utilise de correspondances (et donc plus son score $trust_{degree}$ est élevé) plus sa qualité est assurée. Nous l'observons avec la précision à 1 dès le seuil 0,3.

Pour l'évaluation de la fonction de confiance $trust_{choquet}$, nous réutilisons les scores de qualité des sources utilisées. Les résultats de la fonction $trust_{choquet}$ sont meilleurs sur les seuils bas que ceux de la fonction $trust_{degree}$. On note un très fort rappel et $F - mesure$ pour le seuil 0,1 pour la fonction $trust_{choquet}$. En effet, l'implication des sources pour le candidat permet de contrebalancer l'aspect discriminant des correspondances. En revanche, les deux fonctions ont des résultats identiques sur les seuils hauts. Il est à noter que nos expérimentations n'ont porté que sur la fusion de 3 sources. Des expérimentations impliquant plusieurs sources de qualité variée pourront montrer tout l'impact de la fonction $trust_{choquet}$.

6 Conclusion et Perspectives

Dans cet article, nous avons présenté notre méthode de fusion de plusieurs bases de connaissances. Notre méthode est la première qui travaille avec plus de deux bases. En effet, nous souhaitons extraire de plusieurs sources les éléments consensuels, c'est-à-dire ceux qui sont communs à plusieurs sources. Notre proposition évalue la confiance dans les éléments extraits des sources. Nous avons présenté plusieurs fonctions de confiance. Une évaluation a montré l'intérêt de la fonction $trust_{choquet}$, capable de tenir compte de l'implication locale d'une source dans un élément et de la qualité de cette source. Notre méthode de fusion s'est focalisée uniquement sur la fusion des classes et des individus des BCS. Bien que les aligneurs ne soient pas encore capables de générer des correspondances entre propriétés, nous devons étendre notre méthode de fusion aux candidats représentant les liens entre les classes et les individus.

Références

AMARGER F. (2015). *Vers un système intelligent de capitalisation de connaissances pour l'agriculture durable : construction d'ontologies agricoles par transformation de sources existantes*. PhD thesis, Université de Toulouse 2 le Mirail.

AMARGER F., CHANET J., HAEMMERLÉ O., HERNANDEZ N. & ROUSSEY C. (2014). SKOS sources transformations for ontology engineering : Agronomical taxonomy use case. In *8th Research Conference Metadata and Semantics Research MTSR Karlsruhe, Germany, November, 2014*, p. 314–328.

AMARGER F., CHANET J., HAEMMERLÉ O., HERNANDEZ N. & ROUSSEY C. (2015). Incompatiblity treatment of candidates from several knowledge bases alignments. In *26es Journées Francophones d'Ingénierie des Connaissances, Rennes, juin, 2015*, p. 203–208.

CURÉ O. (2009). Merging expressive ontologies using formal concept analysis. In *On the Move to Meaningful Internet Systems : OTM 2009 Workshops, Vilamoura, Portugal, November, 2009*, volume 5872, p. 49–58.

DRAGISIC Z., ECKERT K., EUZENAT J., FARIA D., FERRARA A., GRANADA R., IVANOVA V., JIMÉNEZ-RUIZ E., KEMPF A. O. & LAMBRIX P. E. A. (2014). Results of the Ontology Alignment Evaluation Initiative 2014. In *9th ISWC workshop on ontology matching (OM), Riva del Garda, Italy, October, 2014*, p. 61–104.

GRABISCH M. & ROUBENS M. (2000). Application of the choquet integral in multicriteria decision making. *Fuzzy Measures and Integrals-Theory and Applications*, p. 348–374.

GUZMÁN-ARENAS A. & CUEVAS A.-D. (2010). Knowledge accumulation through automatic merging of ontologies. *Expert Systems with Applications*, **37**(3), 1991–2005.

JIMÉNEZ-RUIZ E. & GRAU B. C. (2011). Logmap : Logic-based and scalable ontology matching. In *10th International Semantic Web Conference, Bonn, Germany, October, 2011, Proceedings, Part I*, p. 273–288.

NOY N. F. & MUSEN M. A. (2003). The PROMPT suite : interactive tools for ontology merging and mapping. *International Journal of Human-Computer Studies*, **59**(6), 983–1024.

POTTINGER R. A. & BERNSTEIN P. A. (2003). Merging models based on given correspondences. In *29th International Conference on Very large data bases, Berlin, Germany, September, 2003*, p. 862–873.

RAUNICH S. & RAHM E. (2014). Target-driven merging of taxonomies with Atom. *Information Systems*, **42**, 1–14.

SOERGEL D., LAUSER B., LIANG A., FISSEHA F., KEIZER J. & KATZ S. (2004). Reengineering thesauri for new applications : The AGROVOC example. *Journal of Digital Information*, **4**(4).

Réconciliation d'alignements multilingues dans BioPortal

Amina Annane[1], Vincent Emonet[2], Faiçal Azouaou[1] et Clement Jonquet[2, 3]

[1] Ecole nationale Supérieure d'Informatique, Algérie
{a_annane,f_azouaou}@esi.dz
[2]Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier, France
Université de Montpellier & CNRS
{jonquet,vincent.emonet}@lirmm.fr
[3] Center for Biomedical Informatics Research, Stanford University, USA

Résumé : De nos jours, les ontologies sont souvent développées de manière multilingue. Cependant, pour des raisons historiques, dans le domaine biomédical, de nombreuses ontologies ou terminologies ont été traduites d'une langue à une autre ou sont maintenues explicitement dans chaque langue. Cela génère deux ontologies potentiellement alignées mais avec leurs propres spécificités (format, développeurs, versions, etc.). Souvent, il n'existe pas de représentation formelle des liens de traduction reliant les ontologies traduites aux originales et ils ne sont pas accessibles sous forme de *linked data*. Cependant, ces liens sont très importants pour l'interopérabilité et l'intégration de données biomédicales multilingues. Dans cet article, nous présentons les résultats d'une étude de réconciliation des liens de traduction entre ontologies sous forme d'alignements multilingues. Nous avons réconcilié et représenté à l'aide de vocabulaire du web sémantique, plus de 228K mappings entre dix ontologies anglaises hébergées sur le NCBO BioPortal et leurs traductions françaises. Ensuite, nous avons stocké à la fois les ontologies et les mappings sur une version française de la plate-forme, appelée SIFR BioPortal, pour rendre le tout disponible en RDF (données liées). La réconciliation des alignements s'est avérée plus complexe que ce qu'on pourrait penser car les traductions ne sont que rarement l'exacte copie des originales comme nous le discutons.

Mots-clés : Mappings ou alignements multilingues, réconciliations de mappings, web sémantique, données liées, alignement d'ontologies, entrepôt d'ontologies, ontologies biomédicales, BioPortal, interopérabilité et intégration de données, diffusion de données et de connaissances.

1. Introduction

Le domaine biomédical produit de nombreuses ontologies[1]. Cependant la majorité de ces ontologies sont en anglais (Névéol, Grosjean et al. 2014) et même quand des ontologies sont disponibles dans d'autres langues, il y a un manque d'outils pour les utiliser. L'entrepôt de référence d'ontologies biomédicales NCBO BioPortal (http://bioportal.bioontology.org) regroupe plus de 433 ontologies, dont uniquement 6 ne sont pas en anglais, 5 en français et une en espagnol (Jonquet, Emonet et al. 2015). Aussi, le méta-thésaurus d'UMLS (Unified Medical Language System) même s'il couvre 21 langues, 75.1% de ses termes sont en anglais et uniquement 1.82% sont en français (Bollegala, Kontonatsios et al. 2015). Cela représente un frein pour les communautés non anglophones qui produisent et manipulent des données biomédicales dans d'autres langues que l'anglais. Cela les empêche d'exploiter ces ontologies pour la recherche, l'annotation et l'indexation sémantique de leurs données, ainsi que pour l'intégration et l'extraction de connaissance à partir de ces données. En effet, lorsque les ressources biomédicales contiennent des éléments textuels, il est important que la langue de ces ressources soit la même que celles des ontologies qui vont permettre de les exploiter sémantiquement, d'où l'intérêt d'avoir des ontologies multilingues ou de traduire les ontologies d'une langue à une autre (ex : traduction du MeSH par l'INSERM) (Deléger, Merkel et al. 2009; Meilicke, García-Castro et al. 2012). Afin d'assurer l'interopérabilité sémantique, ce qui est un des rôles clés des ontologies, il ne suffit pas de les traduire, mais il

[1] Dans notre contexte le mot ontologie englobe les ontologies et les ressources terminologiques. De la même manière nous parleront de mapping ou d'alignement de façon interchangeable.

faut aussi conserver explicitement les liens entre les objets de l'ontologie traduite et ceux de l'ontologie originale (Buitelaar, Cimiano et al. 2009), d'autant plus si les ontologies continuent à évoluer après leur traduction. (Re)Établir ces liens est l'objet de ce travail que nous appelons réconciliation[2] de mappings multilingues. Nous verrons que même si ce travail ne demande pas du tout la même expertise que la traduction (expertise médicale), ou l'extraction d'alignements (Euzenat and Shvaiko 2013), la tâche devient vite assez difficile si on considère : (i) le nombre d'ontologie et de termes concernés, (ii) leur hétérogénéité, disponibilité, et les façons dont elles ont été traduites et (iii) la variété des champs (e.g., code, identifiant, propriété) qui peuvent être utilisés pour réconcilier les mappings. Ces mappings multilingues, une fois établis et représentés d'une manière formelle, auront de multiples applications (Fu, Brennan et al. 2010). Par exemple, ils peuvent être utilisés pour effectuer une indexation sémantique multilingue de ressources. Ou encore pour intégrer des bases de données biomédicales de langues différentes afin d'étudier des problématiques de recherche telle que la résistance aux maladies, étant donné que les jeux de données spécifiques à une population sont généralement décrits dans la langue de cette population.

Notre travail s'inscrit dans le cadre du projet SIFR (Indexation Sémantique de Ressources biomédicales Francophones – www.lirmm.fr/sifr). Le projet s'intéresse à exploiter les ontologies dans la construction de services d'indexation, de fouille, et de recherche de données pour les ressources biomédicales françaises. Dans ce projet, nous développons un workflow d'indexation basé sur les ontologies (i.e., un annotateur) similaire à celui qui existe pour les ressources anglaises (Jonquet, Shah et al. 2009), mais destiné au français. Pour améliorer le fonctionnement du workflow et relier les ontologies francophones utilisées (hébergées dans une instance locale de BioPortal) à leurs équivalentes anglaises (hébergées sur le NCBO BioPortal ou n'importe où ailleurs), le projet s'intéresse à la réconciliation de mappings multilingues. Cette étude concerne ainsi une dizaine d'ontologies françaises (des traductions) du SIFR BioPortal (http://bioportal.lirmm.fr/ontologies) que nous souhaitons aligner formellement avec les ontologies anglaises originales dans le NCBO BioPortal. Nous nous intéressons à cela seulement pour des ontologies monolingues pour lesquelles il n'existe pas de version multilingues ; en parallèle nous étudions comment à l'avenir, gérer le multilinguisme dans BioPortal (Jonquet, Emonet et al. 2015).

Dans la suite de l'article, nous allons discuter brièvement l'état de l'art du domaine d'alignement des ontologies et la représentation de ressources terminologiques multilingues sur le web (section 2). Ensuite, nous allons étudier la gestion des mappings dans BioPortal et les ontologies traitées dans ce travail (section 3 et 4) ; Expliquer la méthodologie suivie (section 5), analyser les résultats obtenus (section 6), et enfin conclure le travail (section 7).

2. Etat de l'art

Le multilinguisme est l'un des problèmes majeurs d'interopérabilité et d'intégration de données et de connaissances. Ce problème est ressenti de plus en plus à cause de l'explosion des données sur le web sémantique (Buitelaar and Cimiano 2014). Afin de le traiter, différents défis ont été identifiés (Gracia 2012), notamment les alignements multilingues d'ontologies et la représentation des informations lexicales dans les ontologies. Dans la littérature, plusieurs approches ont été proposées pour extraire des alignements multilingues. La première approche adoptée est l'approche manuelle. Il s'agit d'extraire les mappings par les experts humains (Liang and Sini 2006). Cette approche donne des mappings de qualité mais elle est fastidieuse pour les ontologies volumineuses. Par conséquent, les chercheurs se sont retournés vers des approches automatiques. Différentes techniques ont été utilisées : machine learning (Spohr, Hollink et al. 2011), machine translation (Fu, Brennan et al. 2012), extraction de mappings à l'aide de ressources multilingues (Background Knowledge) (Tigrine,

[2] Nous parlons de réconciliation d'alignements car l'information pour formaliser ces alignements existent plus ou moins (comme discuté ci-après) dans les ontologies traduites comparativement aux approches d'extraction de d'alignements qui doivent explicitement détecter des alignements avec des méthodes plus ou moins complexes (structure, syntaxe, etc.).

Bellahsene et al. 2015), etc. Une classification détaillée de ces approches peut être trouvée dans (Trojahn, Fu et al. 2014). Malgré la richesse des travaux dans le domaine de l'alignement d'ontologies, le problème de réconciliation de mappings entre une ontologie et leurs traductions (Euzenat and Shvaiko 2013; Trojahn, Fu et al. 2014) a été ignoré, étant considéré comme une tâche plus facile . Cependant, cette étude montre qu'il n'est pas si facile car les traductions ne suivent pas la même évolution que les ontologies originales ce qui les rend souvent différentes les unes des autres. D'un autre côté et afin de formaliser les liens multilingues entre les ressources terminologiques sur le web (thésaurus, ontologie, etc.), des travaux de recherche ont essayé de définir un modèle capable de représenter les descriptions linguistiques de ces ressources. Le modèle GOLD (General Ontology for Linguistic Description) (Farrar and Langendoen 2003) permet de représenter formellement des concepts linguistiques en utilisant une ontologie owl. Le modèle Lemon (LExicon Model for ONtologies) (McCrae, Spohr et al. 2011) est le modèle le plus complet actuellement pour la publication de ressources lexicales riches sous forme de données liées. En effet, Lemon est le résultat de l'évolution de plusieurs modèles: LMF (Lexical markup framework), LexInfo (Cimiano, Buitelaar et al. 2011), et Linguistic Information Repository (Montiel-Ponsoda, Aguado de Cea et al. 2008). Lemon permet également de représenter les informations lexicales relatives à une ontologie publiée sur le web sémantique. Il a été étendu récemment pour inclure de nouveaux composants comme celui qui traite les traductions (Gracia, Montiel-Ponsoda et al. 2014). Ce qui a donné le modèle OntoLex/Lemon (Bosque-Gil, Gracia et al. 2015). Il est certainement intéressant de disposer de tels modèles pour représenter tous les détails linguistiques d'une ressource, mais il faut réfléchir à leur mise en œuvre aussi. Les modèles riches tels que Lemon sont difficiles à implémenter notamment pour des ontologies très riches et complexes. Par conséquent, les outils qui implémentent ces modèles et facilitent leurs utilisations sont nécessaires pour qu'ils soient adoptés par la communauté (Gracia, Montiel-Ponsoda et al. 2012).

A l'heure actuelle, le domaine biomédical comporte beaucoup d'ontologies monolingues pour lesquelles une traduction a été produite. Parfois par un autre groupe/projet que celui qui a développé l'originale (e.g., MeSH, MedlinePlus, ICD, MEDDRA, et ICPC). La plupart de ces ontologies sont hébergées dans plusieurs plateformes qui veillent à rassembler les ontologies biomédicales et les intégrer dans le but d'offrir des services à la communauté (chacune selon sa vision et son orientation). La gestion du multilinguisme diffère d'une plateforme à l'autre. Le NCBO BioPortal n'est pas multilingue même s'il héberge les ontologies de différentes langues (Jonquet and Musen 2014). Le Métha-thésaurus UMLS, un ensemble de terminologies intégrées manuellement et distribuées d'une manière publique (sauf quelques exceptions) par la bibliothèque NLM (United States National Library of Medicine) (Bodenreider 2004), contient des terminologies (ainsi que leurs alignements) de langues différentes autres que l'anglais (mais ils représentent une minorité, pour le français 1.82% uniquement). Le portail HeTOP (Grosjean, Merabti et al. 2012) offre aussi des termes biomédicaux traduits dans plusieurs langues, notamment le français, il permet également une recherche multilingue mais la plupart de son contenu n'est pas public ou accessible via des services web. Dans ces deux dernières plateformes, l'approche implémentée consiste à intégrer toutes les ontologies en suivant un méta-modèle commun. Cela signifie qu'il existe une abstraction unique pour les concepts de différentes sources (par exemple, les identifiants uniques de concepts d'UMLS (CUI[3])). Ceci est différent de l'approche BioPortal que nous suivons également. Ce dernier ne construit pas un thésaurus global mais garde chaque ontologie séparée et utilise les mappings pour les interconnecter (Ghazvinian, Noy et al. 2009). Une autre différence avec BioPortal, est que ni UMLS ni HeTOP sont construits de manière native avec les technologies du Web sémantique et donc ne proposent pas de représentation sémantique pour rendre les ontologies ou les mappings multilingues disponibles sous format de données liées. En outre, aucune des plateformes citées précédemment n'offre la possibilité de récupérer une ontologie traduite alignée formellement

[3] Concept Unique Identifier : un identifiant unique d'un concept au sein du MetaThesaurus UMLS

avec son origine. Cela représente un frein qui empêche l'exploitation de ces ontologies par des applications multilingues.

L'état de l'art fait ainsi ressortir le besoin (au moins pour le français) de réconcilier les mappings multilingues entre les ontologies traduites et leurs origines ainsi que le besoin de rendre ces mappings disponibles et public sous forme de données liées.

3. Représentation de mappings multilingues dans le BioPortal

Choix des propriétés sémantiques : BioPortal utilise une propriété appartenant aux standards du web sémantique pour étiqueter un mapping entre deux concepts[4]. Par exemple la propriété *exactMatch* du vocabulaire SKOS pour indiquer que deux concepts sont identiques. Les mappings sont sauvegardés dans un format spécifique à la plateforme (e.g., URI1 *skos:exactMatch* URI2), mais cet étiquetage permet de les interroger facilement. En effet, tous les mappings (comme n'importe quel contenu dans le portail) sont disponibles soit à travers un SPARQL endpoint ou via une API REST qui renvoie du JSON-LD. Nous proposons de représenter les mappings multilingues dans BioPortal comme n'importe quel mapping, mais avec des propriétés sémantiques supplémentaires. Des propriétés qui marquent l'aspect linguistique et formalisent le lien de traduction entre deux concepts (l'un est la traduction de l'autre). Par exemple, le concept *Mélanome* de la version française de MeSH sur SIFR BioPortal doit être aligné avec le concept *Melanoma* de la version anglaise de MeSH sur le NCBO BioPortal en précisant qu'en plus du fait que *Mélanome* et *Melanoma* ont la même signification, *Mélanome* est la traduction française de *Melanoma*. Nous pouvons continuer à utiliser les propriétés de SKOS pour représenter la correspondance sémantique entre les concepts, mais pour l'aspect linguistique nous avons besoin d'autres propriétés pour décrire la relation de traduction. Nous proposons d'utiliser les propriétés du modèle Lemon ou GOLD (Jonquet, Emonet et al. 2015). Pour ce travail, nous avons choisi d'utiliser les propriétés de GOLD : la propriété *gold:freeTranslation* pour représenter une traduction exacte (quand les deux concepts ont exactement le même sens, quel que soit leurs libellés), et *gold:translation* pour représenter une traduction moins précise.[5] Ces propriétés répondent à notre besoin (représenter le lien de traduction), sont simples à utiliser, et les ontologies traduites ne contiennent pas les informations liées à la leurs traductions qui nous permettrait d'adopter des vocabulaires plus riches. Par exemple, le concept traduit n'a pas une propriété qui permet de classer sa traduction tel que le vocabulaire Lemon le propose avec *directEquivalent*, *culturalEquivalent* ou *lexicalEquivalent*.

Changements dans l'architecture de BioPortal : Pour pouvoir sauvegarder les mappings multilingues, nous avons dû modifier leur représentation dans BioPortal, notamment pour : (1) permettre d'étiqueter le même mapping avec plusieurs propriétés du web sémantique (cette façon de faire évite de dupliquer les mappings au lieu d'un mapping de traduction et un mapping sémantique, l'étiquetage avec deux propriétés va regrouper les deux dans un seul) ; (2) permettre à BioPortal de stocker des alignements vers des ontologies qui sont soit dans une autre instance de BioPortal (inter-portal), soit vers des ontologies externes qui ne sont dans aucune instance de BioPortal (mappings externes).

4. Les ontologies à aligner

Nous avons traité un ensemble de 20 ontologies (voir Table 1); dix en français et dix en anglais. Ce sont des ontologies très utilisées dans le domaine biomédical pour les deux communautés francophone et anglophone. Par exemple, la Classification Internationale des Maladies (CIM-10) est utilisée dans les hôpitaux pour coder les actes médicaux, le Medical Subject Heading (MeSH) est utilisé pour l'indexation de documents (travaux de la NLM et de

[4] http://www.bioontology.org/wiki/index.php/BioPortal_Mappings

[5] A noter que nous n'utilisons pas la 3ème propriété disponible (gold:literalTranslation) qui identifie des traductions mots à mots et qui en conséquence ne permet pas de représenter une relation entre concepts qui peuvent avoir de multiples labels (nom préféré et synonymes). Voir les définitions de GOLD (http://linguistics-ontology.org/gold) pour plus de détails.

CISMeF). Chaque ontologie française est couplée avec une ontologie anglaise dont elle est issue. Les ontologies anglaises proviennent toutes du Metathesaurus UMLS (version 2015AA) et ont été importées dans le NCBO BioPortal à l'aide de l'outil umls2rdf développé par le NCBO (https://github.com/ncbo/umls2rdf). Les ontologies françaises proviennent soit de l'UMLS (qui contient quelques ontologies directement en français), soit elles ont été produites par le CISMeF du CHU de Rouen qui nous a fourni un fichier OWL pour les inclure dans le SIFR BioPortal. Dans ce deuxième cas, les traductions ont été en général produites ou synthétisées par CISMeF.

TABLE 1 – *Les ontologies traitées dans cette étude (les acronymes sont les identifiants respectifs dans le NCBO ou SIFR BioPortal).*

N°	Ontologie	Acronyme	Version	Format	Source
01	Systematized Nomenclature of MEDicine	SNMI	2015AA	RDF\TTL	UMLS
	Systematized Nomenclature of MEDicine, version française	SNMIFRE	3.5	OWL	CISMeF
02	International Classification of Functioning, Disability and Health	ICF	1.0.2	OWL	UMLS
	Classification Internationale du Fonctionnement, du handicap et de la santé	CIF	2001	OWL	CISMef
03	MedlinePlus Health Topics	MEDLINEPLUS (EN)	2015AA	RDF\TTL	UMLS
	MEDLINEPLUS FR	MEDLINEPLUS (FR)	-	OWL	CISMeF
04	Minimal Standard Terminology of Digestive Endoscopy	MSTDE	2015AA	RDF\TTL	UMLS
	Terminologie minimale standardisée en endoscopie digestive	MTHMSTFRE	2011ab	RDF\TTL	UMLS
05	Semantic Types Ontology	STY (EN)	2015AA	RDF\TTL	UMLS
	Réseau sémantique UMLS	STY (FR)	2014AB	RDF\TTL	CISMeF
06	Medical Subject Headings	MESH	2015AA	RDF\TTL	UMLS
	Medical Subject Headings, version francaise	MSHFRE	2015AA	RDF\TTL	UMLS
07	Medical Dictionary for Regulatory Activities	MEDDRA	2015AA	RDF\TTL	UMLS
	Dictionnaire médical pour les activités règlementaires en matière de médicaments	MDRFRE	2015AA	RDF\TTL	UMLS
08	World Health Organization (WHO) Adverse Reaction Terminology	WHO-ART	2015AA	RDF\TTL	UMLS
	World Health Organization (WHO) Adverse Reaction Terminology, version française	WHO-ARTFRE	1997	OWL	CISMeF
09	International Classification of Diseases, Version 10	ICD10	2015AA	RDF\TTL	UMLS
	Classification Internationale des Maladies, version 10	CIM-10	10	OWL	CISMeF
10	International Classification of Primary Care - 2 PLUS	ICPC2P	2015AA	RDF\TTL	UMLS
	Classification Internationale de Soins Primaires	CISP-2	1998	OWL	CISMeF

5. Méthodologie

5.1. Récupération des triplets à partir des fichiers des ontologies

Les fichiers qui contiennent les ontologies sont au format OWL ou bien RDF/TTL tels que fournis par les BioPortals. Nous avons utilisé l'API Jena pour extraire les triplets RDF à partir des ontologies (sujet-prédicat-objet). Afin de récupérer uniquement les triplets dont nous avons besoin, nous avons filtré selon une propriété précise. En général, cette propriété a pour valeur le code ou l'identifiant que nous souhaitons utiliser pour réconcilier l'alignement. Pour déterminer la propriété adéquate nous avons dû étudier les fichiers contenant les ontologies un par un. La propriété la plus fréquente est *skos:notation*, utilisée pour 12 des 20 ontologies.

D'autres propriétés ont été utilisées telles que *skos:altLabel* (MEDLINEPLUS (FR)) ou *icd:icdCode* (ICF). Dans les cinq cas où la propriété qui contient le code n'existait pas, nous avons extrait le code à partir des URI des concepts. Pour ces derniers cas et afin de ne récupérer que des classes correspondantes à un concept dans l'ontologie, nous avons filtré et conservé uniquement les classes qui ont une propriété *skos:prefLabel*. La Table 2 résume pour chaque ontologie la propriété utilisée pour avoir le code de mapping avec des exemples. Techniquement, nous avons développé une fonction pour chaque ontologie afin d'extraire le code utilisé pour l'alignement. D'éventuels traitements étaient nécessaires tels que l'élimination du type attaché à la valeur, ou l'utilisation d'expression régulière pour isoler la chaine de caractère exacte du code.

TABLE 2– RÉCUPÉRATION DES CODES À UTILISER POUR RÉCONCILIER LES MAPPINGS

Ontology	Code source	Example
SNMI **MTHMSTFRE** **MSTDE** **STY (FR)** **STY (EN)** **MDRFRE** **MEDDRA** **WHO-ARTICD10** **MESH** **MSHFRE** **MEDLINEPLUS (EN)**	Le code interne est affecté à la propriété *skos:notation*	Le concept suivant de l'ontologie MSTDE a pour code interne "**MT200025"**. <http://purl.bioontology.org/ontology/MTHMST/MT200025> a owl:Class ; skos:prefLabel Gastric angioectasia (diagnosis)@eng ; **skos:notation MT200025**^^xsd:string ;
ICF	Le code interne est affecté à la propriété *icd:icdCode*	<owl:Class rdf:ID=s1208> **<icd:icdCode** rdf:datatype=http://www.w3.org/2001/XMLSchema#string >**s1208**</icd:icdCode>
SNMIFRE **CIF** **WHO-ARTFRE** **CIM-10** **CISP-2**	Le code interne est extrait à partir des URI de concepts. Le filtre de triplets se basait sur la propriété *skos:prefLabel*	Le code "**M-40030"** est extrait à partir de l'URI http://chu-rouen.fr/cismef/SNOMED_int.#M-40030 d'un concept de SNMIFRE.
MEDLINEPLUS (FR)	Les concepts n'ont pas de codes internes, nous avons utilisé le CUI pour réconcilier les mappings. Nous avons filter selon la propriété *skos:altLabel*, ensuite nous avons utilisé une expression régulière pour ne récupérer que les CUI et non pas les labels.	<rdf:Description rdf:about=http://chu-rouen.fr/cismef/MedlinePlus#T351> <skos:prefLabel xml:lang=fr>douleur</skos:prefLabel> <skos:altLabel xml:lang=fr>C0008031</skos:altLabel> <skos:altLabel xml:lang=fr>C0030193</skos:altLabel> <skos:altLabel xml:lang=fr>C0030231</skos:altLabel> </rdf:Description>
ICPC2P	Le code interne est affecté à la propriété icpc2p:*icpccode*	<http://purl.bioontology.org/ontology/ICPC2P/ICPCCODE> **A01**^^xsd:string ;

5.2. Réconciliation de mappings

A cette étape nous avons stocké les données extraites à partir des fichiers contenant les ontologies dans une base de données relationnelle (une table par ontologie). La table comporte trois champs : (1) ID, un numéro séquentiel qui identifie chaque enregistrement dans la table ; (2) Code, une chaine de caractère, qui contient le code extrait précédemment et qui peut être un code interne des concepts au niveau de l'ontologie, le CUI ou tout autre critère pertinent de mapping ; (3) URI, une chaine de caractère pour enregistrer l'URI des concepts de l'ontologie étudiée. Ces URI seront nécessaires pour identifier les concepts dans les BioPortals. Chaque triplet récupéré lors de l'étape précédente entraine l'insertion d'un nouvel enregistrement dans la table appropriée. Par exemple le triplet (http://purl.bioontology.org/ontology/MSHFRE/D001542 ; *skos:notation* ; D001542) extrait de la version française de MeSH génère l'enregistrement illustré ci-dessous. Dans la table, il

n'y a pas de contrainte d'unicité pour les deux champs Code et URI. Cela est justifié par le fait que pour un URI donnée il est possible d'avoir plusieurs codes. Nous avons eu ce cas avec la version française de l'ontologie MEDLINEPLUS qui contient 442 concepts ayant plus d'un CUI (identifiant de concept d'UMLS). Par exemple, le concept *minéraux* possède neuf CUI distincts. Nous avons aussi rencontré des cas où le code cible plusieurs URI au sein de la même ontologie (ICPC2P, MEDLINEPLUS FR, CIM-10). Nous allons aborder ces cas plus en détails dans la section 6. Une fois les deux ontologies chargées dans la base de données, nous faisons une jointure sur le champ Code entre les deux tables correspondantes. Comme le code utilisé lors de la jointure n'est pas forcément unique au sein de la même ontologie, le nombre de couples (URI fr, URI en) générés peut être supérieur au nombre de concepts de l'une des deux ontologies (ou les deux). C'est le cas par exemple de l'ontologie CISP2 qui a généré 5063 couples de mappings alors qu'elle n'a que 745 concepts.

TABLE 3–*E*XEMPLE DE CONTENU STOCKÉ DANS UNE TABLE RELATIONNELLE.

Id	Code	URI
1	D001542	http://purl.bioontology.org/ontology/MSHFRE/D001542

5.3. Chargement des mappings dans le SIFR BioPortal

C'est la dernière étape, elle consiste à représenter les alignements produits lors de l'étape précédente d'une manière formelle et permanente dans SIFR BioPortal. Nos mappings ont été systématiquement qualifiés avec une propriété de traduction de GOLD et une propriété d'alignement de SKOS. Au final nous avons utilisé les quatre combinaisons suivantes : (1) *skos:exactMatch/gold:freeTranslation* : utilisée lorsque le concept traduit est exactement le même dans sa version d'origine. C'est généralement le cas lorsque le mapping se base sur une égalité totale entre le code interne du concept français et le code interne du concept anglais. C'est le cas le plus fréquent. (2) *skos:broadMatch/gold:Translation* : pour décrire le lien qui va d'un concept source plus précis que le concept cible. Par exemple le concept *agression par d'autres moyens précisés/établissement collectif* de CIM-10 ayant pour code **Y08.1** n'a pas de concept anglais dans ICD10 avec le même code. Cependant, nous pouvons l'aligner avec le concept *Assault by other specified means* qui a le code **Y08**. (3) *skos:narrowMatch/gold:Translation* : Utilisée pour le cas contraire, c'est lorsque le concept cible est plus précis que le concept source, par exemple le mapping entre CISP2/ICPC2P via le code interne (le code ICPC). Un concept de la version française est aligné avec plusieurs concepts de la version anglaise qui ont le même code ICPC mais différenciés à l'aide d'un suffixe (à cause d'une spécialisation de l'ontologie anglaise après la traduction). Par exemple, le concept tumeur bénigne ayant le code B75 est mappé avec huit concepts anglais plus précis, e.g., (b*enign neoplasm of the blood*, **B75**001), (*benign neoplasm of the lymphatics*, **B75**002), etc. (4) *skos:closeMatch/gold:Translation :* En absence d'un identifiant interne de concepts, nous étions obligés d'utiliser des identifiants moins précis tels que le CUI pour les ontologies qui viennent d'UMLS. Les CUI sont des identifiants au niveau du méta-thésaurus, et non pas au niveau des ontologies sources. Donc, il ne s'agit pas d'une traduction exacte du concept d'une langue à une autre mais plutôt des concepts qui signifient la même chose étant donné qu'on leur a affecté le même CUI.

Tous les alignements produits ont été enregistrés sur le SIFR BioPortal à l'aide d'un script qui utilise l'API web service du SIFR BioPortal (http://data.bioportal.lirmm.fr/documentation). Désormais, pour toutes les ontologies traitées dans ce travail, nous pouvons consulter pour un concept donné, ses mappings multilingues. Le lien ainsi généré permet à l'utilisateur de passer directement du SIFR BioPortal au concept cible dans le NCBO BioPortal afin d'effectuer certain traitements tel que la recherche de ressources anglaises indexées avec ce concept. En plus de l'interface graphique, ces mappings multilingues sont également disponibles via une API web service et un SPARQL endpoint (http://sparql.bioportal.lirmm.fr/test/) ce qui en fait des éléments à part entière du Web de

données et qui les rend facilement réutilisables et exploitables par des applications tierces (via l'utilisation ou non des technologies du web sémantique).

6. Résultats

Notre objectif étant de fournir des alignements pour les versions françaises d'ontologies. Nous exprimerons nos résultats sous forme de pourcentage de concepts de l'ontologie française pour lesquels nous avons pu fournir au moins un mapping de traduction.

Les trois couples d'ontologies (STY FR ; STY EN), (MDRFRE ; MEDDRA), (CIF ;ICF) ont été alignés parfaitement (un pourcentage de 100%) grâce au code interne des concepts. Dans les paragraphes suivants, nous revenons sur l'ensemble des autres couples traités.

MSHFRE/MeSH : Le nombre de concepts de la version anglaise (252242 concept) est dix fois plus grand que le nombre de concepts dans la version française (26142 concept) car cette dernière ne contient que les descripteurs MeSH et pas les concepts supplémentaires[6]. Notre alignement couvre presque la totalité des concepts français avec un pourcentage de 99.79%. Seuls 55 concepts de la version française n'ont pas été alignés car leurs codes n'existent pas dans la version anglaise. De plus, même en essayant de les mapper à l'aide de CUI, nous n'avons pas retrouvé leurs CUI dans le MeSH anglais. Dans le cas d'utilisation de CUI, nous estimons qu'il s'agit sans doute d'erreurs commises par les traducteurs ou des problèmes apparus lors de l'intégration de la nouvelle traduction dans UMLS. En effet, les versions de MeSH devraient être parfaitement alignées car elles viennent les deux d'UMLS. Nous envisageons de contacter l'INSERM et la NLM pour leur indiquer ces résultats.

MTHMSTFRE/MSTDE : Parmi les 1700 concepts de la version française, uniquement deux concepts n'ont pas été alignés car leurs codes n'existent pas dans la version anglaise. Nous avons trouvé un seul concept non mappé dans la version anglaise, il a pour code la valeur *NOCODE*. Cependant ce concept possède deux CUI, qui sont ceux affectés aux concepts français non mappés. Ainsi, pour ces deux concepts français le mapping était par CUI, ce qui a permis d'obtenir 100% d'alignement. Nous estimons que ce cas représente une erreur commise lors de l'intégration de MSTDE dans UMLS, car en principe chaque classe devrait avoir un code. C'est d'ailleurs le cas de la version française MTHMSTFRE.

WHO-ARTFRE/WHO : Le code interne des concepts n'est pas renseigné à travers une propriété dans le fichier contenant l'ontologie française, nous avons dû l'extraire à partir des URI de concepts. Dans la version anglaise WHO-ART, ce code est bien renseigné à l'aide de la propriété *skos:notation*. Nous avons constaté que la version française a subi une modification. En effet, un code de la version anglaise peut référencer plusieurs sous concepts français qui ont le même code de base avec des suffixes différents. Par exemple, le concept anglais qui a le code **1723** référence quatre concepts français ayant pour code : 1723-IT0, **1723-IT1**, **1723-IT2** et **1723-PT**. Pour cela, le nombre de concepts français est supérieur au nombre de concepts anglais (3320 vs 1724). La version française est plus détaillée, ses concepts sont plus précis que ceux de la version anglaise, par conséquent nous avons utilisé les relations *skos:broadMatch*/*gold:translation* pour décrire ces mappings.

MEDLINEPLUS FR/MEDLINEPLUS EN : En raison de l'absence d'un code interne qui distingue les concepts, nous avons utilisé la propriété CUI pour l'alignement de ces deux ontologies. La version française de MedlinePlus contient 795 concepts. Chaque concept a pour propriété un ou plusieurs CUI (442 concepts en ont plus d'un), ce qui a donné 1686 couples (concept, CUI) distincts. La version anglaise contient 1986 concepts distincts, et chaque concept n'a pour propriété qu'un seul CUI. Il est surprenant de noter que des concepts de la version française ont pour propriété des CUI qui n'appartiennent pas à la version anglaise. 123 concepts parmi ces 147 ont pour propriété d'autre CUI appartenant à la version anglaise mais les 24 qui restent n'ont aucun CUI appartenant à la version anglaise (e.g., C0021311, C2362506, etc.). Ces concepts n'existent donc pas ou plus dans la version

[6] Voir http://mesh.inserm.fr/mesh pour des éléments sur la traduction de MeSH.

anglaise. Par conséquent, 24 concepts français n'ont pas été alignés et nous avons obtenu un pourcentage de 97% de concepts français alignés. En essayant de raffiner l'étude, nous avons pris huit de ces concepts qui n'apparaissent pas dans la version anglaise, et nous leur avons appliqué le traitement suivant : (1) Chercher le terme préféré de concept dans l'ontologie française ; (2) Traduire le terme français en anglais manuellement, en utilisant un portail terminologique TermSciences (www.termsciences.fr) ou une autre ressource lexicale (e.g., BabelNet ou même Google translation); (3) Chercher dans la version anglaise, le concept anglais qui a le terme anglais trouvé précédemment comme label et si le concept anglais existe, noter son CUI. Dans 7 cas sur 8, nous avons trouvé le concept anglais qui correspond au concept français (l'origine du concept français) mais avec un CUI différent, comme illustré dans le tableau ci-dessous. Ces résultats nous font penser que ces 24 concepts non mappés sont dû à des erreurs dans le choix des identifiants (CUI) lors du processus de traduction. Nous envisageons de communiquer ces concepts aux traducteurs pour qu'ils détectent les erreurs possibles et éventuellement mettent à jour leur traduction.

TABLE 4 – RÉSULTAT DU TRAITEMENT EFFECTUÉ SUR 8 CONCEPTS

Cui	**PrefLab**	**Cui**	**PrefLabel**
C0156543	Avortement	C0392535	Abortion
C2362506	Fitness et exercice	C1456706	Fitness and Exercise
C0021311	Infections	C3714514	Infections
C1456593	santé mentale et comportement	C1832070	mental health and behavior
C1456620	vivre avec le SIDA	C2963182	Living with HIV/AIDS
C1456571	nutrition des nourrissons et des bébés	/	"nutrition of infants and babies" non trouvé
C2362562	sécurité du patient	C1113679	patient safety
C0002808	Anatomie	C0700276	Anatomy

CISP2/ICPC2P : La version française contient 745 concepts alors que le nombre mapping est de 5141. Cela s'explique par le fait que la version anglaise a subi une modification. Un concept avec un code ICPCCODE a été spécialisé pour générer plusieurs concepts fils (spécialisation des concepts). Par exemple, pour le code ***A01***, dans la version française on trouve un seul concept *douleur générale/de sites multiples* alors que dans la version anglaise ce code référence quatre concepts plus précis (**A01**001; generalised aches), (**A01**004; body pain), etc. Par conséquent, un seul concept de CISP2 génère autant de mappings qu'il y'en a des concepts anglais ayant le même code. Pour cette raison, nous avons étiqueté ces mappings avec les propriétés *skos:narrowMatch/gold:translation*. 59 concepts de l'ontologie française n'ont pas été alignés grâce au code ICPC, en étudiant ces concepts de près, nous avons constaté qu'ils n'ont pas un code ICPC comme le reste des concepts. De plus, ces concepts n'ont aucun CUI comme propriété également. Il semblerait qu'ils ont été ajoutés lors de la traduction, ou supprimés de la version anglaise.

CIM-10/ICD10 : La CIM-10 contient 19853 concepts tandis que sa version anglaise, ICD10, contient 12318 concepts. Pour ce cas aussi nous avons remarqué que la version française a été modifiée. Elle a été enrichie avec de nouveaux concepts résultants d'une spécialisation des concepts originaux. Une jointure selon le code interne des concepts entre les deux ontologies a généré un pourcentage de mapping de 62% (12308 concepts ont été mappés). Nous avons également constaté qu'il existe 6 chapitres dans la version française n'ont pas les mêmes identifiants que leurs correspondants anglais. Par exemple, dans CIM-10 le code d'un chapitre est **(B99)** alors que dans ICD10 le code est **B99-B99.8**. Ces chapitres ont la particularité de ne contenir qu'une seule entrée. Nous avons dû les traiter manuellement vu que la jointure selon le code ne les prenne pas. Tous les mappings discutés précédemment ont été étiquetés avec les propriétés *skos:exactMatch/gold:freeTranslation*. Quant aux concepts générés par une spécialisation (leurs codes n'existent pas dans la version anglaises), nous avons extrait le code de leurs unique direct concept père (les trois premiers digits de leurs codes) et nous les avons mappés avec leurs concepts père anglais. Ces derniers mappings ont été étiquetés avec les propriétés *skos:broadMatch/gold:translation*. Par

exemple, le concept français (*Agression par d'autres moyens précisés /domicile* ; **Y08.0**) a été mappés avec le concept anglais (*Assault by other specified means* ; **Y08).** En suivant ce processus nous avons réduit le nombre de concepts non mappés de 7545 à 40.

SNMIFRE/SNMI : La version française SNMIFRE comporte 106266 concept français, tandis que la version anglaise, contient 109150 concepts, ce qui présente un écart de 2884 concepts en plus dans la version anglaise. Sur la base du code interne, 102093 concepts français ont été mappés, soit 96% de l'ontologie française. Cependant il restait 4173 concepts de la version française sans mapping. En outre, nous n'avons pas pu utiliser non plus la propriété CUI pour ces concepts non mappés, car ils font partie d'un ensemble de 9510 concepts de SNMIFRE qui ne l'ont pas, tandis que les concepts de la version anglaise ont tous cette propriété. Nous n'avons pas trouvé d'autres solutions pour aligner ces 4173 concepts restants.

TABLE 5– RÉCAPITULATIF DES RÉSULTATS OBTENUS

ontologie française	**nombre de concepts**	**ontologie anglaise**	**nombre de concepts**	**Nombre de concepts mappés**	**% de concepts mappés**	**Nombre de Mappings générés**	**Properties (skos ; gold)**
MSHFRE	26142	MeSH	252242	26220	**99.79%**	26220	exactMatch ; freeTranslation
MTHMSTFRE	1700	MSTDE	1699	1700	**100%**	1700	exactMatch ; freeTranslation
STY	133	STY	133	133	**100%**	133	exactMatch ; freeTranslation
MDRFRE	66378	MEDDRA	66378	66378	**100%**	66378	exactMatch ; freeTranslation
MEDLINEPLUS	795	MEDLINEPLUS	2113	771	**97%**	1520	closeMatch ; translation
CIF	1495	ICF	1495	1495	**100%**	1495	exactMatch ; freeTranslation
WHO-ARTFRE	3482	WHO	1724	3482	**100%**	3482	broadMatch ; translation
CISP2	745	ICPC2P	7537	665	**70%**	5063	narrowMatch ; translation
CIM-10	19853	ICD10	12318	19813	**99%**	19813	exactMatch ; freeTranslation 62% broadMatch ; translation 37%
SNMIFRE	106266	SNMI	109150	102093	**96%**	102093	exactMatch ; freeTranslation

7. Conclusion

Dans ce travail nous avons proposé une approche pour la représentation formelle des liens sémantiques reliant une ontologie traduite à son origine. Notre approche consiste à représenter ces liens sous forme de mappings multilingues à l'aide des propriétés sémantiques. Cependant, ce travail ne doit pas être confondu avec l'extraction de mappings multilingues. En effet, nous nous sommes basés dans la majorité des cas sur les codes internes pour réconcilier les liens entre les concepts français et les concepts anglais. Malgré les difficultés que nous avons rencontrées dans certains cas, le lien sémantique entre le concept traduit et son origine existait à travers le code interne des concepts. Notre mission était de le rétablir et de le représenter d'une manière formelle, afin de le rendre disponible pour la communauté. Alors que l'extraction de mappings multilingues consiste à aligner deux ontologies différentes, qui n'ont aucun lien entre elles et qui ne sont pas de la même langue. Cependant, notre approche pour représenter et sauvegarder les mappings peut être utilisée pour représenter les deux types de mappings : réconciliés ou extraits à condition de choisir les propriétés sémantiques adéquates. Dans notre cas, nous avons choisi les propriétés de SKOS et GOLD. Ces propriétés sont complémentaires; en effet, la propriété *gold:translation* ne représente pas la différence entre une relation de traduction *narrow* (vers un concept plus précis que l'autre), une traduction *broad* (vers un concept plus général que l'autre) ou une

traduction *close* (forte relation de similarité) (ces types de traduction sont également identifiés par l'étude de Chen et Chen (S. jiun and H. hua 2012.)), mais en la combinant avec les propriétés de SKOS nous obtenons la description exacte de la relation qui lie les deux concepts. Par exemple, le couple de propriété *skos:narrowMatch*/*gold:translation* décrit une *traduction* de type *narrow*. Un aspect important dans le domaine d'alignement d'ontologies est l'évaluation des mappings résultants (Euzenat and Shvaiko 2013; Trojahn, Fu et al. 2014). Cependant, dans notre cas il s'agit d'une réconciliation de mappings entre concepts basée sur les codes internes de concepts (les codes trouvés à la fois dans l'ontologie originale et l'ontologie traduite) et non pas sur des mesures de similarité (Shvaiko and Euzenat 2013). Par conséquent les mappings obtenus sont systématiquement corrects. Sauf les mappings de MedlinePlus (1% des mappings produits) déduits grâce à la propriété CUI. Vu l'affectation multiples de CUI aux concepts français (section 6), ces mappings par contre doivent être vérifiés. Ainsi que l'ensemble d'anomalies constatées dans certains couples d'ontologies. Nous prévoyons communiquer les résultats aux traducteurs afin de les revoir et les rectifier.

Il est aussi, important de noter que quel que soit la richesse d'une ontologie en termes de langues (2,3 ou même 10), elle ne couvrira jamais toutes les langues. La traduction d'ontologie restera une solution inévitable pour pouvoir exploiter une ontologie dans d'autres langues, autres que celles supportées en natif. Nous espérons que cette étude va sensibiliser les traducteurs concernant l'utilisation des mêmes identifiants au moment de la traduction. Nous pensons que l'idéal est d'utiliser les principes du web sémantique, notamment la réutilisation du même URI comme identifiant unique d'un concept au lieu de créer un nouveau ou l'utilisation d'un alignement *owl:sameAs* qui garantit que les deux concepts sont les mêmes et qu'ils portent en plus la même logique. De plus, dans le processus de création des ontologies multilingues, le défi est d'adopter un standard lexical tel que Lemon au lieu de se contenter de la simple utilisation de la propriété *xmllang* pour spécifier la langue des labels.

Les mappings multilingues produits dans cette étude peuvent avoir de multiples applications, notamment l'intégration de données biomédicales de langues différentes, la recherche et l'indexation sémantiques multilingues. De plus, et dans la continuité du projet SIFR, ces mappings vont être intégrés dans la version française de l'annotateur sémantique (Jonquet, Shah et al. 2009) qui va étendre les annotations directes extraites à partir d'ontologies françaises avec (i) leurs concepts anglais correspondants, (ii) autres ontologies anglaises alignées l'une à l'autre dans le NCBO BioPortal. Nos mappings peuvent aussi être utilisés comme un corpus pour développer des outils de traduction automatique (ou semi-automatique) d'ontologies biomédicales, ce qui peut s'avérer très utile pour les traducteurs d'ontologies. Ils peuvent également servir de *background knowledge* à des approches d'extraction d'alignements entre ontologies différentes. C'est d'ailleurs une perspective que nous regardons actuellement.

Bien entendu, les ontologies évoluent et change à travers le temps (y inclus leur traduction) d'où la nécessité de mettre en œuvre une politique de mis à jour de nos réconciliations (Hartung, Kirsten et al. 2008). Actuellement, nous exécutons le script à nouveau quand une nouvelle version des ontologies traitées est chargée dans le portail, nous supprimons tous les anciens mappings multilingues pour sauvegarder les nouveaux. Ce traitement pourrait être automatisé si le script de réconciliation de mappings multilingues était intégré dans le SIFR BioPortal.

Références

BODENREIDER, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Research 32(suppl 1): D267-D270.

BOLLEGALA, D., G. KONTONATSIOS, ET AL. (2015). A Cross-lingual Similarity Measure for Detecting Biomedical Term Translations. PLoS ONE 10(6).

BOSQUE-GIL, J., J. GRACIA, ET AL. (2015). Applying the OntoLex Model to a Multilingual Terminological Resource. The Semantic Web: ESWC 2015 Satellite Events. F. Gandon, C. Guéret, S. Villataet al, Springer International Publishing. 9341: 283-294.

BUITELAAR, P. AND P. CIMIANO (2014). Towards the Multilingual Semantic Web, Springer.

BUITELAAR, P., P. CIMIANO, ET AL. (2009). Towards linguistically grounded ontologies. The semantic web: research and applications, Springer: 111-125.

CIMIANO, P., P. BUITELAAR, ET AL. (2011). LexInfo: A declarative model for the lexicon-ontology interface. Web Semantics: Science, Services and Agents on the World Wide Web 9(1): 29-51.

DELEGER, L., M. MERKEL, ET AL. (2009). Translating medical terminologies through word alignment in parallel text corpora. Journal of Biomedical Informatics 42(4): 692-701.

EUZENAT, J. AND P. SHVAIKO (2013). Ontology Matching, Springer Berlin Heidelberg.

FARRAR, S. AND D. T. LANGENDOEN (2003). A linguistic ontology for the semantic web. Glot International 7(3): 97-100.

FU, B., R. BRENNAN, ET AL. (2010). Cross-Lingual Ontology Mapping and Its Use on the Multilingual Semantic Web. Multilingual Semantic Web, Raleigh, North Carolina, USA.

FU, B., R. BRENNAN, ET AL. (2012). A configurable translation-based cross-lingual ontology mapping system to adjust mapping outcomes. Web Semantics: Science, Services and Agents on the World Wide Web 15: 15-36.

GHAZVINIAN, A., N. NOY, ET AL. (2009). What Four Million Mappings Can Tell You about Two Hundred Ontologies. 8th International Semantic Web Conference, ISWC'09, Washington DC, USA, Springer.

GRACIA, J. (2012). Cross-lingual ontology matching as a challenge for the Multilingual Semantic Web. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

GRACIA, J., E. MONTIEL-PONSODA, ET AL. (2012). Challenges for the multilingual Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web 11: 63-71.

GRACIA, J., E. MONTIEL-PONSODA, et al. (2014). Enabling language resources to expose translations as linked data on the web. Proc. of 9th Language Resources and Evaluation Conference (LREC 2014), Reykjavik (Iceland), Reykjavik, Iceland, European Language Resources Association.

GROSJEAN, J., T. MERABTI, ET AL. (2012). Multi-terminology cross-lingual model to create the Health Terminology/Ontology Portal. American Medical Informatics Association Annual Symposium.

HARTUNG, M., T. KIRSTEN, ET AL. (2008). Analyzing the evolution of life science ontologies and mappings. Data Integration in the Life Sciences, Springer.

JONQUET, C., V. EMONET, ET AL. (2015). Roadmap for a multilingual BioPortal. MSW4'15: 4th Workshop on the Multilingual Semantic Web, Portoroz, Slovenia.

JONQUET, C. AND M. MUSEN, A. (2014). Gestion du multilinguisme dans un portail d'ontologies: étude de cas pour le NCBO BioPortal. TOTh'14: Terminology and Ontology : Theories and applications Workshop, Bruxelles, Belgium.

JONQUET, C., N. H. SHAH, ET AL. (2009). The Open Biomedical Annotator American Medical Informatics Association Symposium on Translational BioInformatics, AMIA-TBI'09 56-60.

LIANG, A. C. AND M. SINI (2006). Mapping AGROVOC and the Chinese Agricultural Thesaurus: definitions, tools, procedures. New Review of Hypermedia and Multimedia 12(1): 51-62.

MCCRAE, J., D. SPOHR, ET AL. (2011). Linking lexical resources and ontologies on the semantic web with lemon. The semantic web: research and applications, Springer: 245-259.

MEILICKE, C., R. GARCIA-CASTRO, ET AL. (2012). MultiFarm: A benchmark for multilingual ontology matching. Web Semantics: Science, Services and Agents on the World Wide Web 15: 62-68.

MONTIEL-PONSODA, E., G. AGUADO DE CEA, ET AL. (2008). Modelling multilinguality in ontologies.

NEVEOL, A., J. GROSJEAN, ET AL. (2014). Language resources for French in the biomedical domain. 9th International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland.

S. JIUN, C. AND C. H. HUA (2012.). Mapping multilingual lexical semantics for knowledge organization systems. The Electronic Library 30(2): 278–294.

SHVAIKO, P. AND J. EUZENAT (2013). Ontology Matching: State of the Art and Future Challenges. Knowledge and Data Engineering, IEEE Transactions on 25(1): 158-176.

SPOHR, D., L. HOLLINK, ET AL. (2011). A machine learning approach to multilingual and cross-lingual ontology matching. The Semantic Web–ISWC 2011, Springer: 665-680.

TIGRINE, A., Z. BELLAHSENE, ET AL. (2015). Light-Weight Cross-Lingual Ontology Matching with LYAM++. On the Move to Meaningful Internet Systems: OTM 2015 Conferences. C. Debruyne, H. Panetto, R. Meersmanet al, Springer International Publishing. 9415: 527-544.

TROJAHN, C., B. FU, ET AL. (2014). state-of-the-art in multilingual and cross-lingual ontology matching. towards the multilingual semantic web. p. buitelaar and p. cimiano, springer berlin heidelberg: 119-135.

Vers une approche pour la reformulation automatique de requêtes à partir d'alignements complexes

Élodie Thiéblin, Fabien Amarger, Ollivier Haemmerlé, Nathalie Hernandez, Cassia Trojahn

IRIT, Institut de Recheche en Informatique de Toulouse, Toulouse, France
elodie.thieblin@gmail.com,{fabien.amarger, ollivier.haemmerle, nathalie.hernandez, cassia.trojahn}@irit.fr

Résumé : Cet article présente une approche de reformulation automatique de requêtes SPARQL à partir d'un alignement complexe entre deux ontologies. Poser une requête initialement écrite pour une base de connaissances donnée sur une nouvelle base implique de la reformuler à partir du vocabulaire employé dans cette dernière. Un alignement simple permet d'identifier des correspondances entre un élément de la première ontologie et un élément de la deuxième. Ce type d'alignement n'est souvent pas adapté pour exprimer l'ensemble des correspondances pouvant être établies entre deux ontologies. Un alignement complexe permet en effet de déterminer des correspondances plus fines entre 1 ou n éléments de la première ontologie et 1 ou n éléments de la deuxième. Peu d'approches considèrent ce type d'alignements alors qu'il paraît indispensable pour croiser la connaissance provenant des différentes bases de connaissances dont le vocabulaire repose à l'échelle du web de données liés sur des modélisations différentes. Notre approche repose sur des règles de transformation permettant de traduire automatiquement des requêtes SPARQL de type SELECT à partir d'alignements complexes exprimés en EDOAL. Aucun jeu de données ne permettant à notre connaissance de tester ce type d'approche, nous proposons également dans ce papier deux jeux de données que nous avons constitués. Les règles de réécriture ont été validées sur ces jeux de données.
Mots-clés : alignement d'ontologies, alignements complexes, SPARQL, réécriture de requêtes

1 Introduction

Les sources présentes sur le web de données liées sont très hétéroclites. Quand bien même leur format est identique et conforme aux exigences du W3C (RDF, RDFS et OWL), elles restent très hétérogènes par leurs modèles et vocabulaires. Le langage SPARQL est le langage d'interrogation de données RDF. Une requête en SPARQL est propre à la source qu'elle se destine à interroger. Pour cela, la requête SPARQL dépend du modèle ontologique à la base de la source RDF. Si un utilisateur veut poser la même question sur une autre base de connaissances (ou source), il doit adapter la requête SPARQL à cette source.

Combler l'écart d'hétérogénéité terminologique et sémantique entre bases de connaissances devient indispensable pour profiter pleinement du potentiel du web de données liées. L'alignement d'ontologies (Euzenat & Shvaiko, 2007) est une solution à cet enjeu grandissant. Il existe plusieurs types d'alignements : les alignements simples et les alignements complexes. Les alignements simples font correspondre un à un les éléments ontologiques équivalents sémantiquement dans les deux ontologies. Cependant, cette méthode d'alignement ne peut couvrir tous les cas d'utilisation à cause des différences de modèles entre sources ontologiques. Les alignements complexes pallient les faiblesses des alignements simples. Ils étendent les correspondances simples à des correspondances entre constructions complexes d'entités ontologiques.

Les alignements simples sont facilement exploitables lors de la traduction de requêtes

SPARQL. La démarche générale, intégrée à l'Alignment API (David *et al.*, 2011), par exemple, consiste à remplacer l'IRI d'une entité de la requête initiale par l'IRI qui lui correspond dans l'alignement simple, en considérant que la correspondance établie désigne une relation d'équivalence. Cependant, ne considérer que ce type de correspondances ne permet de traduire qu'un ensemble limité de requêtes SPARQL. Dans cet article, nous proposons une approche visant à exploiter les alignements complexes lors de la reformulation de requêtes SPARQL. L'objectif est de proposer un mécanisme s'adaptant au mieux à l'interrogation de sources hétérogènes. Nous souhaitons favoriser le croisement de connaissances et l'utilisation du Web de données liées. Même si peu de systèmes permettent aujourd'hui de générer automatiquement des alignements complexes, établir manuellement ce type de correspondances utilisées en entrée de notre approche est une tâche qui peut s'avérer moins fastidieuse que de reformuler manuellement chacune des requêtes SPARQL potentielles pour un nouvel entrepôt.

Notre approche propose des règles de reformulation pour les requêtes SPARQL de type SELECT à partir des correspondances complexes, impliquant une relation d'équivalence, et exprimables dans la syntaxe EDOAL. Nous proposons également deux jeux de données sur lesquels notre approche a été validée. Le premier correspond à un jeu de données construit pour répondre aux besoins d'experts en agriculture souhaitant trouver sur DBpedia des connaissances complémentaires à celles décrites dans une base de connaissances portant sur les taxons agronomiques. Le second a été conçu en vue d'enrichir le jeu de données de tâche OA4Q de OAEI [1] et porte sur un sous-ensemble du jeu de données en lien avec l'organisation de conférences.

L'article est structuré comme suit. Dans un premier temps, nous définissons les alignements, correspondances et leur formalisation (§2). Ensuite, nous présentons un état de l'art de la reformulation de requêtes SPARQL (§3). Les règles de transformation sur lesquelles repose notre approche (§4) ainsi que les jeux de données sur lesquels elle a été validée sont également présentés (§5).

2 Alignement d'ontologies

Un **alignement** A entre une ontologie source $\mathcal{O}$ et une ontologie cible $\mathcal{O}'$ est un ensemble de correspondances $\{c_1, c_2, ..., c_n\}$ entre $\mathcal{O}$ et $\mathcal{O}'$. A est directionnel et se note $A_{\mathcal{O}\rightarrow\mathcal{O}'}$ (Gillet *et al.*, 2013). Une correspondance c_i est un triplet $\langle e_{\mathcal{O}}, e_{\mathcal{O}'}, r\rangle$:

— si la correspondance est simple, $e_{\mathcal{O}}$ fait référence à une et une seule entité (classe ou propriété) de $\mathcal{O}$ et $e_{\mathcal{O}}$ à une et une seule entité de $\mathcal{O}'$. La correspondance est donc de cardinalité (1:1) ;

— sinon la correspondance est complexe. $e_{\mathcal{O}}$ représente un sous-ensemble d'éléments $\in \mathcal{O}$ et $e_{\mathcal{O}'}$ un sous-ensemble d'éléments $\in \mathcal{O}'$. Les éléments de $e_{\mathcal{O}}$ (resp. $e_{\mathcal{O}'}$) sont reliés entre eux en utilisant un langage formel tel que la Logique du Premier Ordre ou les Logiques de Description. La correspondance est alors de cardinalité (1:n,m:1,m:n).

— r est une relation entre ces deux entités $e_{\mathcal{O}}$ et $e_{\mathcal{O}'}$ telle que l'équivalence ($\equiv$), plus spécifique ($\sqsubseteq$) ou plus général ($\sqsupseteq$).

Une correspondance $\langle e_{\mathcal{O}}, e_{\mathcal{O}'}, r\rangle$ est unique dans un alignement $A_{\mathcal{O}\rightarrow\mathcal{O}'}$. Un alignement est dit complexe si au moins une de ses correspondances est complexe.

1. `http://oaei.ontologymatching.org/2015/conference/index.html`

La syntaxe EDOAL (Euzenat *et al.*, 2007; David *et al.*, 2011) permet d'exprimer les correspondances simples et complexes de cardinalité (1:1), (1:n), (n:1) et (n:m) en précisant la relation entre deux entités ($\sqsupseteq$, $\sqsubseteq$ ou $\equiv$). Les entités e_i et e'_i d'une correspondance complexe c_i sont représentées par des expressions (expressions de classe CE, de relation RE ou de propriété PE) qui peuvent être un ID (ou URI), une construction (composition de plusieurs entités liées par des opérateurs) ou une restriction. Pour une description détaillée des constructeurs définis dans la syntaxe EDOAL, nous renvoyons le lecteur vers (Euzenat *et al.*, 2007).

Nous illustrons par la suite cette syntaxe à partir de correspondances complexes établies entre les ontologies ekaw [2] pour laquelle nous utilisons le préfixe ekaw : <"http://ekaw#">, cmt [3] pour laquelle nous utilisons le préfixe cmt : <"http://cmt#"> et confOf [4] pour laquelle nous utilisons le préfixe confOf : <"http://confOf#">. Ces exemples présentent un fragment d'EDOAL impliquant des relations d'équivalence entre les entités. L'exemple 1 présente une **expression de classe** exprimant que la classe $Chairman$ de l'ontologie cmt est équivalente à l'union des classes $Demo\_Chair$, OC_Chair, PC_Chair, $Session\_Chair$, $Tutorial\_Chair$ et $Workshop\_Chair$ de l'ontologie ekaw. L'exemple 2 établit que la classe $Accepted\_Paper$ de ekaw est équivalente à **l'expression de classe** composée d'éléments ontologiques de cmt restreignant la classe $Paper$ aux individus pour lesquels le domaine de la relation $hasDecision$ est un individu de type $Acceptance$.

Exemple 1

```
<entity1>
  <edoal:Class rdf:about="&cmt;Chairman"/>
</entity1>
<entity2>
<edoal:Class>
  <edoal:or rdf:parseType="Collection">
    <edoal:Class rdf:about="&ekaw;Demo_Chair"/>
    <edoal:Class rdf:about="&ekaw;OC_Chair"/>
    <edoal:Class rdf:about="&ekaw;PC_Chair"/>
    <edoal:Class rdf:about="&ekaw;Session_Chair"/>
    <edoal:Class rdf:about="&ekaw;Tutorial_Chair"/>
    <edoal:Class rdf:about="&ekaw;Workshop_Chair"/>
  </edoal:or>
</entity2>
<measure rdf:datatype="&xsd;float">1.0</measure>
<relation>Equivalence</relation>
```

Exemple 2

```
<entity1>
    <edoal:Class rdf:about="&ekaw;Accepted_Paper"/>
</entity1>
<entity2>
  <edoal:Class>
    <edoal:and rdf:parseType="Collection">
      <edoal:Class rdf:about="&cmt;Paper"/>
        <edoal:AttributeDomainRestriction>
          <edoal:onAttribute>
            <edoal:Relation rdf:about="&cmt;hasDecision"/>
          </edoal:onAttribute>
          <edoal:class>
            <edoal:Class rdf:about="&cmt;Acceptance"/>
          </edoal:class>
        </edoal:AttributeDomainRestriction>
    </edoal:and>
  </edoal:Class>
</entity2>
<measure rdf:datatype="&xsd;float">1.0</measure>
<relation>Equivalence</relation>
```

Dans l'exemple 3, la relation $writtenBy$ de ekaw est équivalente à **l'expression de relation** définissant l'inverse de la relation $writePaper$. L'exemple 4 montre une correspondance entre la classe $Early - Registered\_Participant$ de l'ontologie ekaw et une **expression de classe** définie pour contraindre la classe $Participant$ de l'ontologie confOf à partir d'une **expression de propriété** restreignant la valeur de la propriété $earlyRegistration$ de ses instances à la valeur $true$.

2. `http://oaei.ontologymatching.org/2015/conference/data/ekaw.owl`
3. `http://oaei.ontologymatching.org/2015/conference/data/cmt.owl`
4. `http://oaei.ontologymatching.org/2015/conference/data/confOf.owl`

Exemple 3

```
<entity1>
  <edoal:Relation rdf:about="&ekaw;writtenBy"/>
</entity1>
<entity2>
  <edoal:Relation>
    <edoal:inverse>
      <edoal:Relation rdf:about="&cmt;writePaper"/>
    </edoal:inverse>
  </edoal:Relation>
</entity2>
<measure rdf:datatype="&xsd;float">1.0</measure>
<relation>Equivalence</relation>
```

Exemple 4

```
<entity1>
  <edoal:Class rdf:about="&ekaw;Early-Registered_Participant"/>
</entity1>
<entity2>
  <edoal:Class>
    <edoal:and rdf:parseType="Collection">
      <edoal:Class rdf:about="&confOf;Participant"/>
      <edoal:AttributeValueRestriction>
        <edoal:onAttribute>
          <edoal:Property rdf:about="&confOf;earlyRegistration"/>
        </edoal:onAttribute>
        <edoal:comparator rdf:resource="&edoal;equals"/>
        <edoal:value>
          <edoal:Literal edoal:type="xsd;boolean" edoal:string="true"/>
        </edoal:value>
      </edoal:AttributeValueRestriction>
    </edoal:and>
  </edoal:Class>
</entity2>
<measure rdf:datatype="&xsd;float">1.0</measure>
<relation>Equivalence</relation>
```

3 État de l'art

La réécriture de requêtes SPARQL se fait généralement à partir d'alignements simples. En termes généraux, cela consiste à remplacer l'IRI d'une entité de la requête initiale par l'IRI qui lui correspond dans l'alignement. Cette approche, intégrée à l'Alignment API (David *et al.*, 2011), ne prend pas en compte la relation de correspondance entre les deux entités (généralisation, spécialisation, équivalence). Dans (Euzenat *et al.*, 2008), la combinaison d'extensions de SPARQL pour la réécriture de requêtes SPARQL du type construct, en exploitant les alignements complexes a été proposée. La méthode est appliquée dans un contexte de transformation d'instances entre différents entrepôts de données. Une approche de réécriture ne se restreignant pas à construct et fondée sur des alignements plus expressifs a été proposée par Correndo *et al.* (Correndo *et al.*, 2010), travail dans lequel un ensemble de règles de réécriture de sous-graphes RDF a été défini. Cette approche emploie une formalisation déclarative des alignements entre deux structures RDF. Dans (Correndo & Shadbolt, 2011), un sous-ensemble d'expressions EDOAL sont traduites en leurs patrons de réécriture. Les expressions impliquant les restrictions sur les concepts ou relations, les propriétés différentes de l'égalité et les restrictions sur les occurrences de propriétés ne sont pas prises en compte. Zheng *et al.* (Zheng *et al.*, 2012) présentent une approche de réécriture en se fondant sur la notion de contexte correspondant aux différentes hypothèses sur la façon dont les alignements peuvent être interprétés. Un algorithme de réécriture de triplets RDF prend en compte ces différents contextes et résout de potentiels conflits entre eux. Makris *et al.* (Makris *et al.*, 2010, 2012) présentent le système de réécriture SPARQL-RW qui prend en compte un sous-ensemble de types de correspondances complexes fondés sur une représentation en Logique de Description (i.e., *ClassExpression*, *ObjectPropertyExpression*, *Datatype Property*, et *Individual*). L'algorithme est fondé sur la réécriture de patrons de graphes RDF qui consiste à substituer à un patron initial un patron final en conservant toute variable présente dans le graphe initial. Finalement, Gillet *et al.* (Gillet *et al.*, 2013) proposent une approche de réécriture de patrons de requêtes qui caractérisent des familles de requêtes. L'approche prend en compte des alignements simples et complexes mais n'exploite

pas l'expressivité d'EDOAL.

Dans cet article, nous proposons un ensemble de règles pour la réécriture automatique de requêtes SPARQL, fondé sur des alignements complexes. Contrairement à (Correndo & Shadbolt, 2011), l'approche supporte les restrictions sur les concepts et relations. Cependant, l'approche est limitée aux alignements complexes 1:n et ne prend en compte ni les requêtes sources utilisant les filtres, ni les unions, ni les options SPARQL. Notre approche se fonde sur l'hypothèse selon laquelle les requêtes à reformuler ont pour but de trouver de nouvelles instances répondant à un besoin. Pour cette raison, seuls les éléments de la $\mathcal{T}box$ (informations terminologiques d'une base de connaissances) sont alignés : les classes, les propriétés sur les objets et les propriétés sur les données. La démarche présentée dans cet article cherche à mettre à profit les différentes possibilités d'expressions d'éléments offertes par la syntaxe EDOAL. Comparativement à l'approche de Makris (2010), EDOAL est une alternative permettant la représentation d'alignements plus expressifs. Comme pour les travaux ci-dessous, la génération d'alignements complexes dépasse le cadre de cet article. Très peu d'approches ont éte proposées (Dhamankar *et al.*, 2004; Ritze *et al.*, 2010; Meilicke *et al.*, 2013) et très peu de systèmes gérant les correspondances complexes sont disponibles ou utilisent EDOAL pour les représenter.

4 Approche de reformulation de requêtes SPARQL

Dans nos travaux, nous nous concentrons sur la reformulation de requêtes SPARQL de type SELECT de la forme suivante :

$$R_{\mathcal{O}} \; = \; SELECT \; DISTINCT? \; (\; Var + \; | \; '*' \;) \; WHERE \; \{ \; T_{R_{\mathcal{O}}} \; \}$$

où T_{R_O} correspond à l'ensemble des triplets ou motifs de graphes exprimés dans la requête initiale à partir de l'ontologie $\mathcal{O}$. Un triplet t de cet ensemble est composé d'un sujet s, d'un prédicat p et d'un objet o.

$$\forall t \in T_{R_{\mathcal{O}}}, \quad t = \langle s, p, o \rangle$$

Notre approche vise à produire l'ensemble $T_{R_{O'}}$ contenant les triplets exprimés en fonction de l'ontologie O' à partir de T_{R_O} et de l'alignement complexe $A_{\mathcal{O}\rightarrow\mathcal{O}'}$. Les correspondances complexes (1:n) établissant des relations d'équivalence sont exploitées pour constituer $T_{R_{O'}}$. Les 3-uplets ci-dessous représentent la nature des entités impliquées dans les correspondances à partir des constructeurs EDOAL (constructions et restrictions) (Euzenat *et al.*, 2007) :

$$\langle ClassID, ClassID | ClassConstruction | ClassRestriction, \equiv \rangle$$
$$\langle RelationID, RelationID | RelationConstruction | RelationRestriction, \equiv \rangle$$
$$\langle PropertyID, PropertyID | PropertyConstruction | PropertyRestriction, \equiv \rangle$$

Notre approche analyse la nature des triplets de $T_{R_{\mathcal{O}}}$ en fonction des entités le composant et des correspondances établies dans l'alignement $A_{\mathcal{O}\rightarrow\mathcal{O}'}$. Nous supposons que l'alignement a établi toutes les correspondances nécessaires pour les entités de $T_{R_{\mathcal{O}}}$. Nous définissons des règles qui pour tout triplet de $T_{R_{\mathcal{O}'}}$ génèrent 1 ou n triplets dans $T_{R_{\mathcal{O}'}}$. Deux types de triplets sont considérés : les triplets de types *Classe* et les triplets de types *Relation*. Si un triplet n'est pas identifié comme étant d'un de ces types, il n'est pas traduit mais ajouté dans l'ensemble $T_{R_{\mathcal{O}'}}$.

4.1 Triplets classe

Les triplets classe, notés $T_{R_\mathcal{O}}^{Classe}$, sont de la forme

$$\forall t \in T_{R_\mathcal{O}}^{Classe} t = \langle s, p, o_\mathcal{O} \rangle \quad \text{, où} \begin{cases} s \text{ est une variable} \\ p \text{ est rdf:type} \\ o_\mathcal{O} \text{ est une } \textit{ClassID} \\ \exists < o_\mathcal{O}, o_{\mathcal{O}'}, \equiv > \in A_{\mathcal{O}\to\mathcal{O}'} \end{cases}$$

Un triplet classe est identifié si $o_\mathcal{O}$ est une *ClassID* et si il existe une correspondance impliquant $o_\mathcal{O}$ et une expression de classe $o_{\mathcal{O}'}$. Lors de la transformation d'un triplet classe $t_\mathcal{O}$, le sujet $s_\mathcal{O}$ et le prédicat $p_\mathcal{O}$ restent inchangés. Seul $o_\mathcal{O}$ est traduit suivant l'élément $o_{\mathcal{O}'}$ qui lui correspond dans l'alignement. Les règles de traduction sont fonction de l'expression associée à $o_{\mathcal{O}'}$:

1. *ClassID*: si l'expression de classe $o_{\mathcal{O}'}$ est une *ClassID*, le triplet à rajouter à $T_{R_{\mathcal{O}'}}$ sera

$$\langle s, p, o_{\mathcal{O}'} \rangle$$

2. *ClassConstruction*: Les *ClassExpression* composant $o_{\mathcal{O}'}$ sont annotées comme suit : $o_{\mathcal{O}'}^1$, $o_{\mathcal{O}'}^2$, etc. La transformation du triplet classe dépend de l'opérateur de la relation construction.
   (a) AND : l'intersection mène à l'insertion dans $T_{R_{\mathcal{O}'}}$ de plusieurs triplets ayant le même sujet :
   $$\{\langle s, p, o_{\mathcal{O}'}^1 \rangle \cap \langle s, p, o_{\mathcal{O}'}^2 \rangle \cap ...\}$$
   (b) OR : l'union de plusieurs relations est représentée en SPARQL par une "UNION" entre triplets ayant le même sujet :
   $$\{\langle s, p, o_{\mathcal{O}'}^1 \rangle \cup \langle s, p, o_{\mathcal{O}'}^2 \rangle \cup ...\}$$
   Le tableau 1 donne un exemple de transformation de requêtes à partir de l'exemple 1 de correspondance.
   (c) NOT : consiste à trouver l'ensemble des triplets $\langle s, p, v \rangle$, où v est une variable intermédiaire, duquel on supprime les triplets $\langle s, p, o_{\mathcal{O}'}^1 \rangle$.
   $$\{\langle s, p, v \rangle \text{ MINUS } (\langle s, p, o_{\mathcal{O}'}^1 \rangle) \}$$

3. *ClassRestriction* : les restrictions sur les expressions de classe prennent en compte des expressions de relation ou de propriété, notées $relation(o_f)$ ou $propriete(o_f)$. La transformation du triplet classe dépend de l'opérateur de l'expression correspondante :
   (a) *AttributeTypeRestriction*: cette restriction s'applique à une expression de propriété $o_{\mathcal{O}'}$ consistant à restreindre le type de donnée de cette propriété à un certain $type$. Pour la transformation du triplet de classe t, on ajoute à $T_{R_{\mathcal{O}'}}$ un triplet ayant pour sujet s, pour prédicat l'expression de propriété $propriete(o_{\mathcal{O}'})$ et pour objet une variable intermédiaire v. La restriction se fait sur v à l'aide d'un filtre SPARQL (FILTER) et de la fonction datatype(v).
   $$\{\langle s_i, propriete(o_f), v \rangle \text{ FILTER (datatype}(v) = type \text{) }\}$$

Requête pour cmt	Requête traduite pour ekaw

SELECT ?z WHERE {
?z rdf:type cmt:Chairman.
}
	SELECT ?z WHERE{
	{ ?z rdf:type ekaw:Demo_Chair. }
	UNION { ?z rdf:type ekaw:OC_Chair. }
	UNION { ?z rdf:type ekaw:PC_Chair. }
	UNION { ?z rdf:type ekaw:Session_Chair. }
	UNION { ?z rdf:type ekaw:Tutorial_Chair. }
	UNION { ?z rdf:type ekaw:Workshop_Chair. }
	}

TABLE 1 – Traduction d'un triplet type classe à partir d'une correspondance entre une classe et un expression de classe construite à partir d'une union donnée dans l'exemple 1

Requête pour ekaw	Requête traduite pour cmt

SELECT ?z WHERE {
?z rdf:type ekaw:Accepted_Paper.
}
	SELECT ?z WHERE{
	?z rdf:type cmt:Paper.
	?z cmt:hasDecision ?var_temp.
	?var_temp rdf:type cmt:Acceptance. }

TABLE 2 – Traduction d'un triplet type classe à partir d'une correspondance entre une classe et un expression de classe construite à partir d'une expression de classe donnée dans l'exemple 2

(b) *AttributeDomainRestriction*: cette restriction restreint le codomaine d'une relation issue de l'expression de classe $o_{\mathcal{O}'}$ à une expression de classe $codomaine(o_{\mathcal{O}'})$ elle aussi exprimée dans $o_{\mathcal{O}'}$. Pour traduire t, on ajoute à $T_{R_{\mathcal{O}'}}$ deux triplets. Le premier exprime la relation $relation(o_{\mathcal{O}'})$ entre le sujet s et une variable intermédiaire v. Le deuxième triplet assure que la classe de v soit bien l'expression de classe $codomaine(o_{\mathcal{O}'})$.

$$\{\langle s, relation(o_{\mathcal{O}'}), v\rangle \cap \langle v, rdf : type, codomaine(o_{\mathcal{O}'})\rangle\}$$

Le tableau 2 donne un exemple de transformation de requêtes à partir de l'exemple 2 de correspondance.

(c) *AttributeValueRestriction*: cette restriction s'applique à une expression de relation. Pour limiter les valeurs de l'objet du triplet $relation(o_{\mathcal{O}'})$ on utilise un filtre SPARQL (FILTER) sur la comparaison entre la variable v et la $valeur$ renseignée. Dans l'implémentation actuelle, la valeur renseignée ne peut être qu'un littéral ou une instance. Le comparateur cp correspond à l'un des comparateurs de la syntaxe EDOAL : "=", "<" et ">".

$$\{\langle s, relation(o_{\mathcal{O}'}), v\rangle \text{ FILTER } (v\ cp\ valeur)\},\ cp \in \{=, <, >\}$$

(d) *AttributeOccurenceRestriction*: cette restriction cherche à imposer une contrainte sur le nombre d'occurrences d'une relation ou propriété. Pour compter le nombre d'occurrences en question, on imbrique une sélection en SPARQL (SELECT) du sujet s ainsi que le compte $compte_v$ d'une variable intermédiaire v. Le compte est calculé grâce à la fonction SPARQL COUNT. Ce SELECT imbriqué porte sur le triplet $\langle s, relation(o_{\mathcal{O}'}), v\rangle$. Le compte $compte_v$ est ensuite comparé à la valeur de la restric-

tion val_{rest} grâce au comparateur cp dans un filtre SPARQL (FILTER).

$$\begin{array}{l}\{\{\text{SELECT } s \text{ (COUNT}(v) \text{ AS } compte_v) \text{ WHERE}\\ \{\langle s, relation(O_{\mathcal{O}'}), X\rangle\}\\ \text{GROUP BY } s.\ \}\\ \text{FILTER } (compte_v\ cp\ val_{rest})\},\\ cp \in \{=, <, >\}\end{array}$$

4.2 Triplets relation

Les triplets relation, notés $T^{Relation}_{R_{\mathcal{O}}}$, sont de la forme

$$\forall t \in T^{Relation}_{R_{\mathcal{O}}} t = \langle s, p_{\mathcal{O}}, o\rangle \quad \text{, où } \begin{cases} s \text{ une variable} \\ p_{\mathcal{O}} = \text{ une RelationId ou PropertyId} \\ o \text{ une variable ou un littéral} \\ \exists < p_{\mathcal{O}}, p_{\mathcal{O}'}, \equiv > \in A_{\mathcal{O}\to\mathcal{O}'} \end{cases}$$

Dans les triplets relation, $p_{\mathcal{O}}$ est une *PropertyId* et $p_{\mathcal{O}'}$ une expression de relation ou propriété. Lors de la transformation d'un triplet, la nature de l'expression est considérée :

1. *RelationId* ou *PropertyId*: le triplet suivant est inséré dans $T_{R_{\mathcal{O}'}}$

$$\langle s, p_{\mathcal{O}'}, o_i\rangle$$

2. *RelationConstruction* ou *PropertyConstruction*: la transformation dépend de l'opérateur de la construction. Les opérateurs suivis d'une * ne sont valables que pour les RelationConstructions. Les RelationExpressions ou PropertyExpressions composant $p_{\mathcal{O}'}$ seront indicées comme suit: $p^1_{\mathcal{O}'}$, $p^2_{\mathcal{O}'}$,...
   (a) AND : cette construction peut avoir un ou plusieurs composants *RelationExpression*. Les triplets générés seront :
   $$\{\langle s, p^1_{\mathcal{O}'}, o\rangle \cap \langle s, p^2_{\mathcal{O}'}, o\rangle \cap ...\}$$
   (b) OR : cette construction peut avoir un ou plusieurs composants *RelationExpression*. Les triplets générés seront :
   $$\{\langle s, p^1_{\mathcal{O}'}, o\rangle \cup \langle s, p^2_{\mathcal{O}'}, o\rangle \cup ...\}$$
   (c) NOT : la négation d'un triplet relation correspond à l'ensemble des triplets $\langle s, v, o\rangle$, où v est une variable, duquel on retranche les triplets $\langle s, p^1_{\mathcal{O}'}, o\rangle$. Les triplets générés seront :
   $$\{\langle s, v, o\rangle \text{ MINUS } (\ \langle s, p^1_{\mathcal{O}'}, o\rangle\ )\ \}$$
   (d) COMPOSE : une composition de relations est une chaîne de relations. Des variables intermédiaires v_1, v_2, ... sont introduites pour compléter la chaîne de relations entre le sujet s et l'objet o. On considère qu'une imbrication de composition ou qu'une imbrication d'une négation au sein d'une composition serait un problème de modélisation de

l'alignement. Cette *RelationConstruction* peut avoir un ou plusieurs composants *RelationExpression*. Dans la transformation, p est traduit par un ensemble de triplets de la forme :

$$\{\langle s, p^1_{\mathcal{O}'}, v_1\rangle \cap \langle v_1, p^2_{\mathcal{O}'}, v_2\rangle \cap ... \cap \langle v_{n-1}, p^n_{\mathcal{O}'}, o\rangle\}$$

(e) INVERSE* : inverser une relation revient à intervertir son sujet et son objet. Cette construction ne peut avoir qu'un composant *RelationExpression*. Le triplet généré sera :

$$\langle o, p^1_{\mathcal{O}'}, s\rangle$$

Le tableau 3 donne un exemple de transformation de requêtes à partir de la correspondance donnée en exemple 3 ;

(f) REFLEXIVE* : une relation réflexive est une relation entre le sujet s et lui-même. Sa représentation SPARQL est donc un triplet (sujet, relation, sujet). Elle ne peut avoir qu'un composant *RelationExpression*. Le triplet généré sera :

$$\langle s, p^1_{\mathcal{O}'}, s\rangle$$

(g) SYMMETRIC* : la symétrie d'une relation est l'union entre une relation et son inverse. Cette construction ne peut avoir qu'un composant *RelationExpression*. Les triplets générés seront :

$$\{\langle s, p^1_{\mathcal{O}'}, o\rangle \cup \langle o, p^1_{\mathcal{O}'}, s\rangle\}$$

3. *RelationDomainRestriction* ou *PropertyDomainRestriction*: pour restreindre le domaine d'une relation à une expression de classe, est rajoutée à $T_{R_{\mathcal{O}'}}$:

$$\{\langle s, p^1_{\mathcal{O}'}, o\rangle \cap \langle s, rdf : type, domain(p^1_{\mathcal{O}'})\rangle\}$$

4. *RelationCoDomainRestriction*: pour restreindre le codomaine d'une relation à une expression de classe, est rajoutée à $T_{R_{\mathcal{O}'}}$:

$$\{\langle s, p^1_{\mathcal{O}'}, o\rangle \cap \langle o, rdf : type, domain(p^1_{\mathcal{O}'})\rangle\}$$

5. *PropertyTypeRestriction*: pour restreindre le type de donnée d'une propriété, on utilise un filtre SPARQL (FILTER) agrémenté de la fonction "datatype(objet)" pour assurer l'égalité entre le $type$ souhaité et le type de donnée de l'objet : datatype(o).

$$\{\langle s, p^1_{\mathcal{O}'}, o\rangle \text{ FILTER (datatype}(o) = type)\}$$

6. *PropertyValueRestriction*: la restriction d'une propriété sur la valeur de son objet peut être représentée par un filtre SPARQL (FILTER) sur la comparaison entre l'objet de la propriété et la valeur renseignée. Dans la version actuelle, la valeur renseignée ne peut être qu'un littéral. Le comparateur cp correspond à l'un des comparateurs de la syntaxe EDOAL : "=", "<" et ">". Le tableau 4 donne un exemple de transformation d'une requête à partir d'une correspondance impliquant une expression de classe, elle-même définie à partir d'une expression de propriété restreignant la valeur, donnée en exemple 4.

$$\{\langle s, p^1_{\mathcal{O}'}, o_i\rangle \text{ FILTER } (o\ cp\ valeur)\},\ cp \in \{=, <, >\}$$

Requête pour ekaw	Requête pour cmt

SELECT ?z WHERE {
?paper :writtenBy ?author.
}
	SELECT ?z WHERE{
	?author cmt:writePaper ?paper.
	}

TABLE 3 – Traduction d'un triplet relation à partir de la correspondance de l'exemple 3

Requête pour ekaw	Requête pour confOf

SELECT ?z WHERE{
?z rdf:type ekaw:Early-Registered_Participant.
}
	SELECT ?z WHERE{
	?z rdf:type confOf:Participant.
	?z confOf:earlyRegistration ?var_temp.
	FILTER(?var_temp="true"^^ xsd:boolean).
	}

TABLE 4 – Traduction d'un triplet type classe impliquant une expression de classe définie à partir d'une expression de relation donnée dans l'exemple 4

5 Validation

À notre connaissance, aucun jeu de données intégrant deux bases de connaissances, l'alignement complexe entre les deux ontologies concernées et des requêtes SPARQL écrites pour les deux bases, n'est disponible. Seuls des jeux de données composés d'alignements simples existent dans le cadre de la tâche oa4qa [5] de la campagne OAEI. En reprenant le principe de cette tâche, nous avons constitué manuellement deux nouveaux jeux de données afin de valider notre approche.

Bases de connaissances et requêtes SPARQL. Le premier jeu de données a été construit lors d'un projet cherchant à accéder à des connaissances en lien avec une taxonomie des plantes. Pour répondre à ce besoin, les bases de connaissances Agronomic Taxon [6] et DBpedia ont été considérées. Il était plus précisément question de retrouver la connaissance suivante :

— $qa1$: les taxons de type espèce
— $qa2$: les taxons ayant pour rang taxinomique supérieur un taxon de type famille
— $qa3$: les taxons de rang taxinomique règne
— $qa4$: les taxons de rang taxinomique ordre
— $qa5$: les taxons de rang taxinomique genre

Pour constituer le jeu de données, les requêtes SPARQL de référence correspondant à ces besoins décrits en langage naturel ont été écrites manuellement pour chacune des deux bases de connaissances. L'objectif identique à celui suivi dans la tâche oa4qa est de pouvoir vérifier le fait qu'une requête reformulée automatiquement renvoie les mêmes résultats que la requête de référence.

Nous avons suivi la même démarche pour constituer le deuxième jeu de données. Il vise à interroger un sous-ensemble d'ontologies du jeu de données OAEI 2015 portant sur l'organisa-

5. http://oaei.ontologymatching.org/2015/
6. http://ontology.irstea.fr/AgronomicTaxon

Requête initiale posée sur ekaw	Requête de référence pour confOf	Requête générée pour confOf

SELECT ?person WHERE {
?person :authorOf ?paper.
?paper a :Paper.
?person rdf:type :Early-Registered_Participant.
}
	SELECT ?person WHERE{
	?person :earlyRegistration true.
	?person :writes ?papier.
	?papier a :Paper.
	}
		SELECT ?person WHERE {
		?person :writes ?paper.
		?paper a :Paper.
		?person rdf:type :Participant.
		?person :earlyRegistration ?variable_temp0.
		FILTER(?variable_temp0 = "true"^^xsd:boolean).
		}

TABLE 5 – Exemple de requête initiale, de requête de référence et de requête générée

tion de conférences [7]. Nous avons plus précisément considéré trois ontologies (cmt, confOf et ekaw). Nous avons défini en langage naturel les besoins suivants, à savoir retrouver :

— $qb1$: les reviewers de papiers acceptés
— $qb2$: les auteurs de soumissions longues
— $qb3$: les chairmen qui ont soumis un papier
— $qc1$: les participants qui se sont inscrits tôt et qui sont auteurs d'un papier soumis
— $qc2$: les participants qui se sont inscrits tard et qui ont écrit un poster

Les trois ontologies ont été peuplées avec des instances répondant à ces besoins. Parallèlement, les besoins ont été traduits en requêtes SPARQL écrites spécifiquement pour chacune des bases de connaissances considérées.

Les requêtes sur ces deux jeux de données sont disponibles en ligne [8].

Alignements complexes. Pour le jeu de données Agronomic Taxon et DBpedia, 10 correspondances complexes (et 1 simple) ont été produites manuellement. Pour le jeu de données Conférences, 8 correspondances simples et 6 correspondances complexes ont été construites manuellement. Les alignements sont disponibles en ligne [9].

Discussion. Le tableau 5 montre pour le besoin $qc1$, la requête SPARQL initialement posé sur $ekaw$, la requête de référence considérée pour la base $confOf$ et la requête générée par notre approche pour cette base à partir de l'exemple 4 et de la correspondance simple $ekaw : authorOf \equiv confOf : writes$. Bien que la requête de référence et la requête générée soient syntaxiquement différentes, leur exécution renvoie le même ensemble de résultats. Ceci est le cas pour l'ensemble des requêtes considérées. Les requêtes générées sont disponibles en ligne [10].

6 Conclusion et perspectives

Dans ce papier, nous avons présenté une approche qui exploite un alignement complexe entre deux ontologies pour réécrire des requêtes SPARQL formulées pour une ontologie source en

7. http://oaei.ontologymatching.org/2015/conference/index.html
8. https://www.irit.fr/recherches/MELODI/telechargements/requetes.zip
9. https://www.irit.fr/recherches/MELODI/telechargements/alignements.zip
10. https://www.irit.fr/recherches/MELODI/telechargements/requetesgenerees.zip

requêtes SPARQL formulées pour une ontologie cible. Une première validation de l'approche a permis de mettre en place deux jeux de données qui devront à très court terme être étoffés.

Les pistes d'amélioration sont nombreuses car le mécanisme de traduction se limite à des requêtes formatées, uniquement composées de triplets dont le sujet est une variable. Une autre piste de recherche serait la prise en compte des correspondances (n:m) d'un alignement complexe. Dans notre implémentation, la relation d'une correspondance entre deux entités (généralisation, spécialisation ou équivalence) n'est pas prise en compte. Une recherche sur leur signification lors d'une traduction de requête SPARQL peut être intéressante. Une autre extension possible consiste à considérer la transitivité des relations. Nous pouvons également améliorer certains points du mécanisme de traduction, notamment le fait que la *PropertyValueRestriction* est appliquée seulement sur des littéraux, ou que l'*AttributeValueRestriction* porte seulement sur les instances et les littéraux.

Références

CORRENDO G., SALVADORES M., MILLARD I., GLASER H. & SHADBOLT N. (2010). SPARQL Query Rewriting for Implementing Data Integration over Linked Data. In *1st International Workshop on Data Semantics (DataSem 2010)*.

CORRENDO G. & SHADBOLT N. (2011). Translating expressive ontology mappings into rewriting rules to implement query rewriting. In *6th Workshop on Ontology Matching*.

DAVID J., EUZENAT J., SCHARFFE F. & TROJAHN C. (2011). The Alignment API 4.0. *Semantic Web*, **2**(1), 3–10.

DHAMANKAR R., LEE Y., DOAN A., HALEVY A. & DOMINGOS P. (2004). imap: Discovering complex semantic matches between database schemas. In *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, SIGMOD '04, p. 383–394.

EUZENAT J., POLLERES A. & SCHARFFE F. (2008). Processing ontology alignments with sparql. In *International Conference on Complex, Intelligent and Software Intensive Systems*, p. 913–917.

EUZENAT J., SCHARFFE F. & ZIMMERMANN A. (2007). *Expressive alignment language and implementation*. Rapport interne, INRIA.

EUZENAT J. & SHVAIKO P. (2007). *Ontology Matching*. Berlin, Heidelberg: Springer-Verlag.

GILLET P., TROJAHN C., HAEMMERLÉ O. & PRADEL C. (2013). Complex correspondences for query patterns rewriting. In *Proceedings of the 8th International Workshop on Ontology Matching*.

MAKRIS, GIODALSIS B. C. (2010). Ontology mapping and sparql rewriting for querying federated rdf data sources.

MAKRIS K., BIKAKIS N., GIOLDASIS N. & CHRISTODOULAKIS S. (2012). SPARQL-RW: transparent query access over mapped RDF data sources. In *15th International Conference on Extending Database Technology*, p. 610–613: ACM.

MAKRIS K., GIOLDASIS N., BIKAKIS N. & CHRISTODOULAKIS S. (2010). Ontology mapping and SPARQL rewriting for querying federated RDF data sources. In *2010 Conference on On the Move to Meaningful Internet Systems*, p. 1108–1117.

MEILICKE C., NOESSNER J. & STUCKENSCHMIDT H. (2013). Towards joint inference for complex ontology matching. In *Late-Breaking Developments in the Field of Artificial Intelligence*.

RITZE D., VÖLKER J., MEILICKE C. & SVÁB-ZAMAZAL O. (2010). Linguistic analysis for complex ontology matching. In *5th Workshop on Ontology Matching*.

ZHENG X., MADNICK S. E. & LI X. (2012). SPARQL Query Mediation over RDF Data Sources with Disparate Contexts. In *WWW Workshop on Linked Data on the Web*.

Rôle d'une base de connaissance dans SemIoTics, un système autonome contrôlant un appartement connecté

Nicolas Seydoux[1,2,3], Khalil Drira[2,3], Nathalie Hernandez[1], Thierry Monteil[2,3]

[1] IRIT Maison de la Recherche, Univ. Toulouse Jean Jaurès,
5 allées Antonio Machado, F-31000 Toulouse
{nseydoux,hernande}@irit.fr

[2] CNRS, LAAS, 7 avenue du Colonel Roche,
F-31400 Toulouse, France
{nseydoux,khalil,monteil}@laas.fr

[3] Univ de Toulouse, INSA, LAAS, F-31400, Toulouse, France

Abstract : L'Internet des Objets représente une réalité de plus en plus concrète au fur et à mesure que se déploient de larges réseaux d'objets connectés. Ceux-ci ouvrent de larges perspectives d'applications, mais rencontrent des difficultés en terme d'interopérabilité, de configuration ou de passage à l'échelle. Ces problématiques peuvent être traitées par le recours aux principes du web de données liées, d'où l'émergence d'ontologies dédiées aux applications de l'IoT, comme IoT-O, une ontologie pour l'IoT.Par ailleurs, une description enrichie des systèmes permet d'envisager leur configuration autonome : on parle alors d'autonomic computing. Ce papier présente SemIoTics, un système autonome reposant sur des bases de connaissance pour la gestion d'un appartement connecté. Nous présentons tout d'abord une vision générique d'une architecture de réseaux d'objets connectés qui permet de guider une analyse des travaux à l'interface du web sémantique et de l'IoT. Nous décrivons ensuite les deux bases de connaissances spécialisant IoT-O sur lesquelles s'appuie SemIoTics, et leur relation avec le dispositif expérimental. Enfin, la structure de ce système autonome de domotique est présenté en détails, et mis en relation avec l'architecture identifiée dans l'état de l'art.
Mots-clés : Web sémantique, Internet des Objets, Autonomic computing, Domotique

1 Introduction

L'Internet des Objets, Internet of Things (IoT) en anglais, désigne des réseaux d'objets connectés communiquant les uns avec les autres pour étendre leurs fonctionnalités Gubbi *et al.* (2013). Dans un premier temps plus axés sur les capteurs connectés Barnaghi *et al.* (2012), les réseaux d'objets tendent à inclure des objets de plus en plus variés : objets collectant l'énergie, actionneurs (objets qui agissent sur le monde)... Le développement important de l'IoT permet d'envisager des applications dans des domaines nombreux : ville intelligente, agriculture, domotique, usine intelligente, télésanté, etc. L'hétérogénéité des domaines d'application tend à poser des problèmes d'interoperabilité entre les solutions, dites verticales, développées dans une approche orientée silos. Cette problématique d'interopérabilité se pose à deux niveaux : l'interoperabilité architecturale et l'interopérabilité sémantique, décrites dans Gyrard *et al.* (2015). L'interopérabilité architecturale est composée de l'interoperabilité technique et organisationnelle, et les efforts de standardisation actuels tentent de la résoudre. L'interopébilité sémantique se pose sur le sens associé aux interactions entre les objets (appels de services, découverte de fonctionnalités...) ainsi qu'aux données qu'ils échangent. C'est sur cette forme d'interopérabilité que se concentre la communauté du web sémantique, en appliquant des principes et des technologies rendant les données plus compréhensibles par les systèmes.

De plus, résoudre ce problème d'interopérabilité amène à une autre problématique majeure de l'IoT : la complexité des systèmes. En effet, comme le soulignent Zanella *et al.* (2014), Barnaghi *et al.* (2012) ou Foteinos *et al.* (2013), la grande hétérogénéité des composants d'un système d'objets connectés en fait des entités complexes à gérer, surtout à grande échelle. Plus les interactions se multiplient, plus les technologies mises en jeu sont variées, et plus la gestion du système par des opérateurs humains est difficile et coûteuse. À partir de ce constat, qui englobe mais ne se limite pas au domaine des objets connectés, Kephart & Chess (2003) propose un nouveau paradigme, l'autonomic computing, qui vise à rendre possible l'auto-configuration ou l'auto-réparation des systèmes. Les comportements autonomes sont guidés par des politiques de haut niveau définies par les opérateurs humains, et par la connaissance que le système possède sur ses composants. Nous proposons dans ce papier de représenter cette connaissance à partir des formalismes du web sémantique, ce qui permet de traiter à la fois la problématique de la complexité du système par l'autonomic computing et la problématique de l'interoperabilité.

Pour appuyer ce double apport du web sémantique à l'IoT, nous présentons SemIoTics, un système autonome de contrôle d'objets connectés guidé par une Base de connaissances (BC). Nous démontrons que SemIoTics, élaboré de façon générique décorrélée de tout domaine d'application particulier, est applicable au domaine spécifique de la domotique. Nous étudions pour cela un cas d'utilisation reposant sur le contrôle d'un appartement connecté. Le reste de ce papier est structuré comme suit. Tout d'abord, LMU-N (Lower, Middle and Upper Node), une structuration architecturale générique pour l'IoT, est proposée pour structurer la présentation de l'état de l'art sur l'intégration des principes et des technologies du web sémantique dans les réseaux d'objets connectés. Ensuite, après des rappels concernant l'ontologie IoT-O proposée dans Seydoux *et al.* (2015), nous décrivons la connaissance exploitée par le système qui reposent sur des modules spécialisant IoT-O. Enfin, l'architecture de SemIoTics est présentée, et mise en relation avec LMU-N.

2 LMU-N : une architecture pour classifier les contributions du web sémantique à l'IoT

2.1 Motivations et caractéristiques de la classification guidée par LMU-N

La notion d'IoT est fondée sur la notion de réseau d'objets connectés, un graphe où les sommets (que l'on appellera ici noeuds) sont les objets, et les arcs les interactions qui existent entre eux. Cette vision des objets connectés sous forme de noeuds amène une unification de deux composants fondamentaux de l'IoT De *et al.* (2011) : l'objet et le service. En effet, un noeud du réseau d'objets peut être vu comme un service : son interface est connue de ses noeuds voisins, et son implémentation sous-jacente peut être associée à un objet matériel comme elle peut être purement logicielle. Un arc est établi entre deux noeuds quand des données ou des services sont échangées entre les deux, ou plus simplement quand au moins un des deux noeuds a conscience de l'existence de l'autre et peut accéder à son interface. Les arcs sont donc orientés, depuis le noeud capable d'initier l'interaction vers l'autre.

Les noeuds d'un IoT sont très divers en terme de puissance de calculs, de proximité avec le monde physique ou de capacités de communication. La puissance de calcul représente la capacité de l'objet à appliquer des traitements complexes à des données en masse, ainsi que la complexité des types de données supportés. L'étendue des capacités de communications est mesurée à l'aulne du nombre de protocoles que l'objet utilise pour communiquer, de la bande

passante dont il dispose, et de sa disponibilité. La proximité avec le monde physique se caractérise par le nombre d'arcs à parcourir depuis le noeud considéré pour effectuer une action ayant directement prise avec l'environnement, que ce soit pour interagir avec un noeud effectuant une mesure (par exemple, accéder à une donnée issue d'un capteur de température) ou pour actionner un noeud permettant d'effectuer une modification sur le comportement d'un objet (par exemple lors de l'allumage ou le réglage d'un radiateur). À partir de ces trois paramètres (puissance, proximité avec le monde physique, capacités de communication), trois catégories homogènes de noeuds se dégagent, comme illustré dans la table 1 :

- Les **noeuds de haut niveau**, typiquement serveurs ou ordinateurs de bureau ou portables, représentent les noeuds faiblement contraints de l'architecture de Zanella *et al.* (2014), ou les noeuds de traitement de données dans Liu *et al.* (2015).

- Les **noeuds médians**, qui assurent la connexion entre les noeuds de haut niveau et les noeuds de bas niveau, sont souvent associés à la notion de gateway comme dans Ben-Alaya *et al.* (2015) et Desai *et al.* (2015). Ces noeuds n'ont pas pour rôle principal de traiter les données, mais de transformer les informations depuis les noeuds de haut niveau pour leur utilisation par les noeuds de bas niveau.

- Les **noeuds de bas niveau** représentent tous les objets ou programmes connectés qui sont le plus directement en prise avec le monde physique, mais très contraints en terme de puissance, de consommation et de communication. Ces objets sont par définition présents dans toutes les architectures d'IoT.

LMU-N (Lower, Middle and Upper Node) est le nom de cette architecture générique en trois niveaux de noeuds. LMU-N recouvre donc les architectures proposées dans Zanella *et al.* (2014), Liu *et al.* (2015), Desai *et al.* (2015), Ben-Alaya *et al.* (2015),Nikoli *et al.* (2011), Mrissa *et al.* (2015)Pour un noeud d'un niveau donné, on constate dans les architectures analysées qu'il aura des contacts privilégiés avec un petit nombre de noeuds de niveau supérieur, et avec de multiples noeuds du niveau inférieur. De plus, les connexions entre noeuds de même niveau vont croissant avec la "hauteur" de ce niveau. Un schéma représentant cette architecture générique est représenté sur la figure 2. En unifiant l'objet au service, et donc le réel au virtuel, la notion de noeud permet aussi l'abstraction des objets réels élémentaires en objets virtuels plus élaborés. Cette abstraction, proposée dans Foteinos *et al.* (2013), permet de composer des objets et services réels en des noeuds composites. À l'inverse, un noeud complexe peut être abstrait et décomposé en un ensemble de noeuds plus simples, chacun exposant une sous-partie cohérentes de ses fonctionnalités.

Toutefois, identifier ces trois niveaux d'architectures n'est pas suffisant pour situer précisément les façons dont le web sémantique permet de lever les verrous de l'IoT. Dans la suite de cette section, les flux opérationnels dans LMU-N ainsi que les connaissances qu'ils mettent en jeu sont caractérisés et mis en relation avec les travaux de l'état de l'art.

2.2 Classification des apports du web sémantique à l'IoT sous forme de flux dans LMU-N

La notion de noeud de réseau s'articule avec la notion de flux opérationnel (workflow en anglais) partiellement définie dans Poslad *et al.* (2015). Ces flux sont de deux types : montants (up-

Type de noeud	Puissance de calcul	Capacités de communication	Proximité avec le monde physique	Exemples
Noeud de haut niveau	Importante	Étendues (bande passante importante, protocoles supportés variés...)	Très faible	Serveurs, smartphones, ordinateurs
Noeud médian	Moyenne à faible	Étendues (bande passante importante, protocoles supportés variés...)	Faible	Gateway, box domestique
Noeud de bas niveau	Faible à très faible	Limitées (peu de bande passante, support de protocoles contraints uniquement)	Très forte	Capteur, actionneur

Figure 1: Caractérisation des types de noeuds d'un réseau d'objets connectés

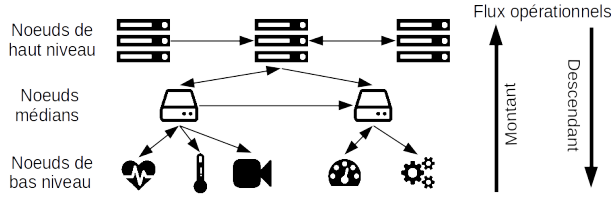


Figure 2: Représentation de LMU-N, de ses trois niveaux de noeuds et ses flux opérationnels

stream) si le noeud émetteur est de niveau inférieur au noeud récepteur, et descendant (downstream) dans le cas contraire. Cette distinction est rendue nécessaire par les contraintes liées aux différents niveau de noeuds. Ces flux opérationnels mettent en oeuvre des connaissances de nature variée, décrites par des ontologies. Celles-ci sont de deux types : les ontologies spécifiques à un domaine d'application, et les ontologies core-domaine dédiées à la description des objets et des services qu'ils offrent d'une manière décorrélée de l'application. Un grand nombre d'ontologies de ces deux types sont répertoriés dans le LOV4IoT[1], initiative visant à favoriser la réutilisation d'ontologies déjà existantes.

- **Enrichissement montant et traduction descendante** : Ce flux opérationnel porte sur les données manipulées par le système, dont le niveau d'enrichissement doit être modulé selon le niveau du noeud la manipulant. La remontée d'information permet d'enrichir une donnée pour la transformer en information, voire en connaissance, par l'annotation sémantique, comme dans Sheth *et al.* (2008), Le-phuoc & Hauswirth (2009) ou Poslad *et al.* (2015). À l'inverse, un noeud de haut niveau peut transformer une information pour l'exprimer sous une forme compréhensible par un noeud de niveau inférieur, tout en devant garantir son intégrité. Les ontologies mises en jeu sont principalement des ontologies de domaine en lien avec les domaines visés par les systèmes, ainsi que des ontologies dédiées à l'IoT permettant de décrire comment les données sont liées au système, comme dans Sheth *et al.* (2008).

- **Exposition montante et découverte descendante** : Pour que des noeuds distants puissent accéder à son interface, tout noeud doit procéder à une étape d'exposition de celle-ci. Dans le cas des noeuds de bas niveau, il est possible qu'un noeud intermédiaire assure

[1] `http://www.sensormeasurement.appspot.com/?p=ontologies#home`

leur exposition, en général un noeud médian. Le noeud intermédiaire est dans ce cas dépositaire de connaissance concernant le noeud de bas niveau, qu'il contextualise et rend accessible, agissant comme un proxy Nikoli *et al.* (2011). La découverte est l'opération complémentaire à l'exposition, puisqu'elle consiste à prendre conscience des interfaces exposées. La découverte dynamique de noeuds est conditionnée par leur description compréhensible, permise par un vocabulaire partagé, comme souligné dans Barnaghi *et al.* (2012). Ce flux opérationnel porte sur les connaissances représentant les objets constituants le système, et s'appuie donc principalement sur des ontologies de core-domaine de l'IoT. Les connaissances échangées décrivent les capacités des noeuds, les services qu'ils offrent, leur état de fonctionnement, les moyens par lesquels ils peuvent être contactés... L'ensemble de ces connaissances sur les noeuds permet, la découverte passée, une sélection guidée par un besoin applicatif, comme dans Perera *et al.* (2014) où des objets sont sélectionnés à partir d'une description sémantique.

- **Notification montante et contrôle descendant** : dans un réseau d'objets connectés, les objets les plus complexes commandent les objets les plus simples. Par exemple, les applications sur smartphones allument les lampes ou collectent les informations de capteur. Pour éviter la scrutation active, certaines architectures incluent un mécanisme de notification qui permet à un noeud de bas niveau de faire passer une information à un noeud de haut niveau, comme prévu dans le standard oneM2M[2] par exemple. Ce flux opérationnel peut être associé au premier flux pour enrichir les notifications et traduire les commandes.

Ces trois types de flux opérationnels sont génériques à l'IoT, et permettent de caractériser les apports des principes et des techniques du web sémantique à l'IoT. Ces flux peuvent être composés pour construire des applications complexes : dans la section suivante, nous établissons un parallèle entre ces flux et une boucle générique d'autonomic computing.

2.3 Décomposition d'une boucle autonomique basée sur une base de connaissance dans LMU-N

L'autonomic computing est un paradigme de programmation qui vise à limiter l'intervention humaine dans le fonctionnement des systèmes complexes Kephart & Chess (2003). Des propriétés de configuration autonome Chatzigiannakis *et al.* (2012) ou de gestion autonome Vlacheas *et al.* (2013) par exemple facilitent la mise en oeuvre à grande échelle de systèmes d'objets connectés hétérogènes. Le rôle de l'opérateur humain est transformé : il fixe des objectifs de haut niveau que le système va répercuter de manière cohérente sur ses différents composants. La gestion autonome d'un système s'appuie sur une boucle générique, dite boucle MAPE-K Kephart & Chess (2003) : Monitoring, Analysis, Planning, Execution, le tout lié par une BC (Knowledge). La boucle de gestion autonomique est une approche tout à fait générique, abstraite de tout domaine d'application. Elle peut donc être utilisée pour guider le comportement d'un réseau d'objets connectés. Dans ce cas, les différentes étapes d'une boucle MAPE-K peuvent être rapportées à des flux dans l'architecture LMU-N :

- L'**observation** (Monitoring) est un **flux montant de notification et d'enrichissement**

[2] `http://onem2m.org/`

d'information. Les données brutes produites par les capteurs sont enrichies et stockées dans la BC de l'agent, et entrent ainsi dans son processus de prise de décision.

- L'étape d'**analyse** (Analysis) s'appuie sur du filtrage de données, du raisonnement et sur les connaissances de l'agent pour donner un sens aux signaux qu'il a observés. Elle est effectuée dans un noeud médian ou de haut niveau.
- L'étape de **planification** peut avoir lieu dans des noeuds médians ou des noeuds de haut niveau. Elle consiste en une **découverte** de noeuds de bas niveau et en une prise de décision répondant à l'analyse de la situation, guidée par les **politiques de haut niveau** définies par l'administrateur. L'ensemble des connaissances nécessaires à cette étape est stocké dans la BC de l'agent, et celle-ci est enrichie par les déductions issues de l'analyse ainsi que par les décisions découlant de la planification.
- L'étape d'**exécution** correspond à un **flux descendant de traduction et de contrôle**. À partir de ses connaissances de haut niveau, l'agent va émettre des données qui instancient ces connaissances dans un format sémantiquement dégradé, mais cohérent dans le contexte de leur interprétation par les objets qui en sont la cible.

La notion de BC utilisée dans Kephart & Chess (2003) est à prendre au sens large : il s'agit d'un ensemble de règles et d'informations qui guident le comportement de l'agent. Nous proposons de représenter ces connaissances à l'aide des formalismes du web sémantique. C'est sous cette forme que les connaissances de l'agent sont exprimées dans SemIoTics : il dispose d'un ensemble de règle et d'individus qui instancient des ontologies. L'agent tire parti de l'ouverture des données, du partage de vocabulaires, ainsi que de la maturité des formalismes du W3C qui permettent de mettre en oeuvre du raisonnement.

3 Spécialisation de IoT-O pour représenter la connaissance nécessaire à la gestion intelligente d'un bâtiment

SemIoTics implémente une boucle MAPE-K qui s'appuie sur une BC exprimée dans les formalismes du W3C. Cette connaissance est décrite à l'aide de IoT-O et de modules qui l'étendent, et caractérise l'appartement connecté géré par SemIoTics.

3.1 IoT-O, une ontologie de core-domain pour l'IoT

IoT-O[3] est une ontologie décrivant des connaissances génériques à tous les domaines de l'IoT. Sa structuration initiale a été proposé dans Ben-Alaya *et al.* (2015), et une version enrichie, structurée par des patrons de conception a été présentée dans Seydoux *et al.* (2015). Sa structure modulaire régie par des patrons, ainsi que son import de nombreuses ressources existantes en font une ontologie de core-domain pour l'IoT-O plus facile à réutiliser et à étendre que d'autres ontologies de l'IoT. la figure 3 donne une vision de haut niveau de l'architecture de IoT-O.

Cette figure souligne la modularité de IoT-O, construite sur 5 axes : Observation, Action, Service, Énergie et Cycle de vie. Les noms des modules créés dans la construction de IoT-O sont en rouge surligné, et le nom de ceux qui ont demandé des modifications (alignement

[3] `http://www.irit.fr/recherches/MELODI/ontologies/IoT-O`

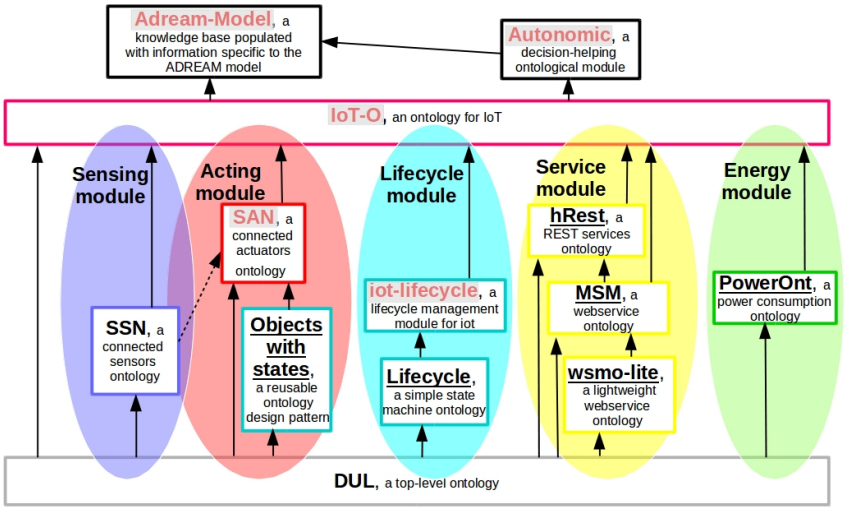


Figure 3: Architecture de haut niveau de IoT-O

avec DUL par exemple) est souligné. En tant qu'ontologie de core-domaine, IoT-O est conçue pour être spécialisée par des modules dédiés à une application particulière. Le reste de cette section est dédié à la description de Autonomic, module étendant IoT-O pour représenter la connaissance nécessaire à la mise en place d'un agent autonome, ainsi que de Adream-Model, module spécialisant lui aussi IoT-O pour représenter la connaissance propre à notre système.

3.2 Autonomic : un module spécialisant IoT-O pour mettre en place un agent autonome

Le module Autonomic[4] est une BC qui étend IoT-O pour permettre d'orienter les choix d'un agent autonome par rapport aux politiques de haut niveau fixées par l'utilisateur et aux données collectées par les objets. Ce module contient la connaissance nécessaire à la mise en place d'une boucle MAPE-K basée sur une BC, décrite dans la section 2.3.

Le module Autonomic définit la notion de *ConstrainedProperty*, une propriété de l'environnement (température, luminosité...) que l'utilisateur veut maintenir entre des bornes qu'il fixe, en spécialisant la propriété de IoT-O importée de SSN *ssn:Property*, représentant toute propriété dont le système a connaissance. Cette même classe *ssn:Property* est étendue par les classes *AboveMaxValueProperty* et *BelowMaxValueProperty*, qui servent à classifier les propriétés dont une observation permet de déterminer qu'elles ne respectent pas une contrainte.

Cette classification s'appuie sur une étape de raisonnement. Les observations sont liées aux objets qui les ont produites, ce qui permet à l'agent de prendre en compte leurs caractéristiques dans son traitement. Le module Autonomic spécialise la notion d'*iot-o:Impact*, pour indiquer si les actions qu'un objet peut effectuer sont de nature à faire augmenter (*PositiveImpact*) ou diminuer (*NegativeImpact*) la valeur de la propriété impactée. De plus, un *iot-o:Impact* peut être relatif (*RelativeImpact*) si le système peut en graduer l'effet (par exemple, fixer une valeur de consigne) ou absolu (*AbsoluteImpact*) dans le cas d'une interaction binaire de type on/off.

[4] `https://www.irit.fr/recherches/MELODI/ontologies/Autonomic`

3.3 Adream-Model : un module représentant la connaissance propre à ADREAM

3.3.1 Description de l'appartement connecté

Le système SemIoTics a été appliqué sur le bâtiment ADREAM[5], qui comporte un ensemble d'objets connectés, capteurs et actionneurs. Le cas d'utilisation présenté dans ce papier s'appuie sur un sous-ensemble de ceux-ci, organisés en deux systèmes séparés : la gestion de la température et la gestion de la luminosité. Pour gérer la température, l'appartement dispose d'un ventilateur, d'un chauffage électrique, ainsi que d'un capteur de température. La gestion de la luminosité repose sur un capteur de luminosité et une lampe.

Les objets sont connectés à une plateforme open-source, OM2M, implémentant le standard oneM2M, ce qui assure leur interopérabilité. La plateforme associe aux objets une interface REST qui permet d'y accéder de manière uniforme. Toutefois, dans sa version actuelle, oneM2M n'assure que l'interoperabilité architecturale. L'interopérabilité sémantique nécessite donc l'utilisation des technologies du web sémantique pour enrichir les données et les rendre interopérables et compréhensibles afin d'en garantir la consistance d'une application à l'autre, problème évoqué dans Corcho & García-Castro (2010).

La plateforme s'exécute sur une gateway supportant différents protocoles : Phidget, ethernet, sigfox, LoRa, enOcean. Pour établir un parallèle avec LMU-N, les objets connectés sont des noeuds de bas niveau : chacun ne maîtrise qu'un protocole de communication, et n'est dédié qu'à une tâche simple. La gateway sur laquelle s'exécute la plateforme est un noeud médian : elle permet la communication selon divers protocoles, et est suffisamment puissante pour exécuter le code de la plateforme, mais ne dispose que de capacités de traitement et de stockage limitées. La plateforme n'intègre pas de noeud de haut niveau, l'agent autonome est donc exécuté sur une machine tierce communiquant avec le noeud médian intégré à l'appartement.

3.3.2 Caractérisation des noeuds avec le module Adream-Model

Le module Adream-Model[6] spécialise IoT-O en proposant un ensemble de classes, de relations et d'individus permettant la description des objets connectés accessibles dans l'appartement, ainsi que des services qui en ouvrent l'accès. Un schéma illustrant l'instanciation du sous-système gérant la température est présenté dans la figure 4. L'élément au coeur du système est la propriété avec laquelle il interagit, ici *qudt:LuminousIntensity*. Cette propriété est mesurée par un capteur et impactée par deux actionneurs, l'ensemble de ces objets ayant des interfaces de *msm:Service* constituées d'*msm:Operations*. Celles-ci sont associées aux impacts que les objets peuvent avoir sur la propriété (augmentation ou diminution), et l'accès à ces opérations est déterminé par une machine à état modélisant le fonctionnement de l'objet. Adream-Model réutilise aussi une partie du vocabulaire du module Autonomic pour décrire les objets de Adream, pour que l'agent puisse prendre des décisions les concernant.

[5] `https://www.laas.fr/public/en/adream`

[6] `https://www.irit.fr/recherches/MELODI/ontologies/Adream-Model`

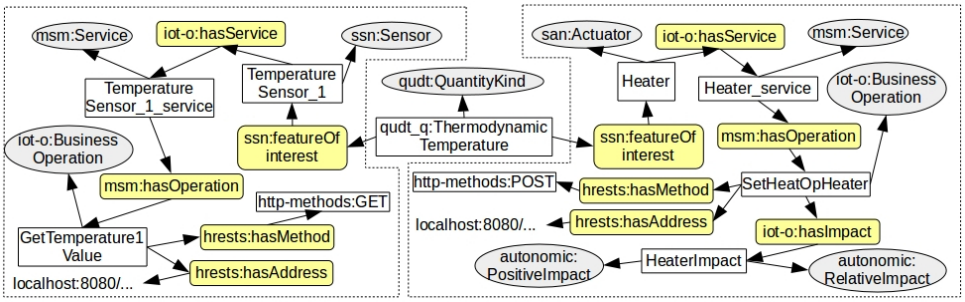


Figure 4: Extrait de la base de connaissance décrivant la gestion de température dans ADREAM

4 Mise en oeuvre : contrôle autonomique d'un appartement

4.1 SemIoTics, un système guidé par les décisions de l'utilisateur

Les décisions de l'agent autonome sont guidées par les contraintes imposées par l'utilisateur. Celles-ci sont vues comme des objectifs de haut niveau qu'il incombe au système d'atteindre. L'utilisateur exprime ses contraintes à travers une interface graphique, celles-ci sont traduites en connaissances et directement incorporées à la BC. Le coeur de SemIoTics reposant sur des traitements lourds (raisonnement, code Java) et demandant des capacités de stockage importantes pour la BC, le système s'exécute donc dans un noeud de haut niveau de LMU-N.

La BC de SemIoTics n'est pas exclusive à l'agent, elle expose une interface qui permet à un tiers de la consulter ou d'y ajouter des données. Cette configuration rend l'interaction avec les objets dynamiques : la gateway ou un opérateur humain peuvent déclarer des modifications dans les noeuds de bas niveau, rendant possible la découverte d'un nouveau noeud, la mise à jour de caractéristiques d'un noeud connu ou la suppression d'un noeud devenu inaccessible. Ce dynamisme permet la modification du comportement de l'agent en temps réel par des connaissances injectées par un opérateur. Cette BC est donc le point d'aboutissement des flux opérationnels d'exposition, et les flux de découverte par les noeuds de haut niveau s'appuient sur les informations qui y sont stockées. Pour les noeuds médians, la découverte de noeuds de bas niveau doit être assurée par un autre moyen, comme OM2M pour SemIoTics.

4.2 Monitoring : D'une donnée mesurée par un capteur à une connaissance

La collecte des données depuis les capteurs est assurée en premier lieu par la plateforme OM2M. Celle-ci offre un accès unifié aux données, et assure toujours un accès à la donnée la plus récente mesurée par le capteur. L'agent autonome de SemIoTics interroge la plateforme grâce aux descriptions des services associés aux capteurs contenues dans la BC (voir figure 4). À partir de cette description, la donnée brute retournée par la plateforme, contenant une date, une valeur et une unité est transformée en une *ssn:ObservationValue*. Cet enrichissement permet d'expliciter la connaissance implicite dont le système est en possession : le lien est établi avec le capteur à son origine, ainsi qu'avec la propriété sur laquelle porte l'observation. Ce processus est un flux opérationnel montant d'enrichissement décrit par LMU-N.

4.3 Analysis : Abstraction de l'observation en évènements porteurs de sens

L'étape d'analyse consiste à déterminer si les éléments observés à l'étape de monitoring demandent une réaction de la part de l'agent. Celle-ci peut donc entièrement se dérouler dans le noeud de haut niveau dans lequel SemIoTics s'exécute. Un premier filtrage est effectué sur les observations pour ne tenir compte que de la plus récentes portant sur chaque propriété contrainte par l'utilisateur. Ensuite, une étape de raisonnement permet de déterminer si une propriété est violée par l'observation en question. Ce raisonnement s'appuie sur une règle définie à l'aide du moteur de règle de Jena[7] embarqué dans SemIoTics. Ce raisonnement amène à la création de symptômes représentant une abstraction des observations. Si une action de l'agent est nécessaire, il entre en phase de planification.

4.4 Planification : Détermination des actions à effectuer par l'agent

Pendant l'étape de planification, l'agent autonome va déterminer sa réaction face au contexte analysé précédemment. Cette étape, comme l'étape d'analyse, est guidée par la connaissance stockée dans la BC de SemIoTics, et s'exécute donc dans un noeud de haut niveau. La propriété à corriger, ainsi que le type de correction à apporter (augmentation, diminution) sont établies dans l'étape d'analyse. L'agent autonome liste tous les objets agissant sur cette propriété, et sélectionne dans cette liste les objets pouvant avoir l'impact demandé.

À cette étape, si aucun objet n'est approprié pour répondre au besoin du système, la contrainte est déclarée insatisfiable (*autonomic:UnsatisfiableProperty*). Ce marquage est réévalué dès que l'ensemble d'objets connu par l'agent est modifié. Tant qu'une contrainte sur une propriété est estimée insatisfiable par le système, celui-ci ne tient pas compte des observations portant sur cette propriété dans la phase de monitoring. Cette mesure permet d'éviter de relancer des raisonnements inutiles tant qu'une action significative n'a pas été entreprise.

Pour chaque objet, l'agent va lister les services qu'il offre et effectuer un nouveau tri par rapport à leur description dans IoT-O. Par exemple, deux types de services sont identifiés : les services liés à la fonctionnalité spécifique de l'objet, *iot-o:BusinessService*, définis par opposition au service de management liés à l'entretien de l'objet, comme la consultation du niveau de batterie, la mise à jour du firmware, etc. De même, pour les services sélectionnés, l'agent découvre les opérations qu'ils exposent, leurs effets, et leur utilité potentielle concernant l'état de l'objet. C'est là que la modélisation des actionneurs sous forme de machine à état à l'aide de IoT-Lifecycle[8] intervient : les actions sont associés à des états, et déclenchent des transitions. Ainsi, pour augmenter la luminosité, l'agent ne considérera pas qu'allumer une lampe déjà en marche est une action valide. Après ces différentes étapes de filtrage, l'agent sélectionne une opération parmi celles qui ont été retenues, et ajoute dans la base de données une *san:Actuation* représentant l'appel à celle-ci. L'étape d'exécution s'appuie ensuite sur cette *san:Actuation*.

Dans notre cas, l'ensemble des filtrages effectués par l'agent est guidé par la nécessité. Dans un cas plus général où plusieurs objets pourraient convenir, cette sélection pourrait être enrichie par des expressions de politiques guidées par la qualité de service Chaocan Xiang *et al.* (2015), la consommation énergétique, ou toute autre politique de haut niveau exprimée comme dans Perera *et al.* (2014).

[7] `https://jena.apache.org/documentation/inference/index.html`

[8] `https://www.irit.fr/recherches/MELODI/ontologies/IoT-Lifecycle`

4.5 Execution : De l'abstraction à la commande

L'étape d'exécution met en place des décisions actées par l'agent pendant la planification, c'est donc un flux descendant de contrôle. Ce flux de contrôle est doublé d'un flux de traduction : le modèle de données utilisé dans le noeud de haut niveau prenant la décision peut ne pas être compris par le noeud ciblé. La traduction peut être vue comme une "dégradation sémantique" de la connaissance vers une donnée qui assure la consistance du sens de la donné dans le contexte où elle est utilisée par rapport au sens de la connaissance dont elle dérive. La traduction a lieu en deux étapes : le noeud de haut niveau transmet d'abord au noeud médian les connaissances nécessaire au contrôle. Le noeud médian construit ensuite une requête adaptée au noeud de bas niveau de destination, en tenant compte du format de données ou du canal de communication qu'il supporte. La commande est alors envoyée au noeud de bas niveau, qui met à jour son état en conséquence et permet au système d'agir sur le monde physique. La modification de la propriété impactée est mesurée par des capteurs, qui injecteront ces mesures dans le système à la prochaine étape de monitoring, amenant à un contrôle de l'environnement en boucle fermée.

5 Conclusion et travaux futurs

La contribution principale de ce papier est la proposition de SemIoTics, un système autonome de gestion d'un ensemble d'objets connectés s'appuyant sur une BC. Cette proposition est accompagnée d'une structuration architecturale générique, LMU-N, qui permet de caractériser les contributions du web sémantique à l'IoT. SemIoTics vise à surmonter certains verrous de l'IoT : les contraintes matérielles des objets, le manque d'interopérabilité, et la complexité du système dûe à l'hétérogénéité et au passage à l'échelle. La notion d'objet contraint est prise en compte dans LMU-N, ce qui permet sa prise en compte dans des flux opérationnels adaptés. SemIoTics s'appuie sur une plateforme standard, OM2M, pour apporter l'interopérabilité architecturale, et sur les principes du web sémantique pour construire l'interopérabilité sémantique. Enfin, la complexité du système est considérée par l'utilisation de la boucle de gestion autonomique, combinée à une BC : les données y sont enrichies à la fois par des connaissances spécifiques au domaine d'application, et par des connaissances sur le système lui-même. Ces dernières sont décrites par des modules spécialisant IoT-O, une ontologie de core-domaine pour l'IoT.

Dans le futur, un open data permettra de consulter les données issues du bâtiment instrumenté du LAAS, ADREAM. Cet open data comprendra une interface permettant d'accéder à certaines fonctionnalités de SemIoTics, et d'un endpoint permettant une exploration des bases de connaissances de l'agent autonome. Les sources de données de SemIoTics s'étendront des objets de l'appartement (quelques dizaines d'objets) aux capteurs de ADREAM tout entier (plus de 2400 capteurs). Cette évolution amènera des problématiques de passage à l'échelle demandant une gestion des données en flux, mais aussi de qualité des données (capteurs défectueux, incohérences de la base de données) ou d'abstraction qui reposent sur de l'aggrégation de données issues de capteurs redondants ou complémentaires.

References

BARNAGHI P., WANG W., HENSON C. & TAYLOR K. (2012). Semantics for the Internet of Thingss: early progress and back to the future. In *International Journal on Semantic Web and Information Systems*, volume 8, p. 1–21.

BEN-ALAYA M., MEDJIAH S., MONTEIL T. & DRIRA K. (2015). Toward semantic interoperability in oneM2M architecture. *IEEE Communications Magazine*, **53**(12), 35–41.

CHAOCAN XIANG, PANLONG YANG, XUANGOU WU, HONG HE & SHUCHENG XIAO (2015). QoS-based service selection with lightweight description for large-scale service-oriented internet of things. In *Tsinghua Science and Technology*, volume 20, p. 336–347: Tsinghua University Press (TUP).

CHATZIGIANNAKIS I., HASEMANN H., KARNSTEDT M., KLEINE O., KRÖLLER A., LEGGIERI M., PFISTERER D., RÖMER K. & TRUONG C. (2012). True Self-Configuration for the IoT. In *3rd International Conference on the Internet of Things (IOT)*.

CORCHO O. & GARCÍA-CASTRO R. (2010). Five challenges for the Semantic Sensor Web. *Semantic Web*, **1**(1), 121–125.

DE S., BARNAGHI P., BAUER M. & MEISSNER S. (2011). Service modelling for the Internet of Things. In *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*, p. 949–955.

DESAI P., SHETH A. & ANANTHARAM P. (2015). Semantic Gateway as a Service architecture for IoT Interoperability. In *Kno.e.sis Publications*.

FOTEINOS V., KELAIDONIS D., POULIOS G., VLACHEAS P., STAVROULAKI V. & DEMESTICHAS P. (2013). Cognitive management for the internet of things: A framework for enabling autonomous applications. *IEEE Vehicular Technology Magazine*, **8**(4), 90–99.

GUBBI J., BUYYA R., MARUSIC S. & PALANISWAMI M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, **29**(7), 1645.

GYRARD A., SERRANO M. & ATEMEZING G. A. (2015). Semantic web methodologies, best practices and ontology engineering applied to Internet of Things. In *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)*, p. 412–417: IEEE.

KEPHART J. & CHESS D. (2003). The vision of autonomic computing. *Computer*, **36**(1), 41–50.

LE-PHUOC D. & HAUSWIRTH M. (2009). Linked open data in sensor data mashups. In *Proceedings of the 2nd International Workshop on Semantic Sensor Networks (SSN09)*, volume 522, p. 1–16.

LIU J., LI Y., CHEN M., DONG W. & JIN D. (2015). Software-defined internet of things for smart urban sensing. *IEEE Communications Magazine*, **53**(9), 55–63.

MRISSA M., MEDINI L., JAMONT J.-P., LE SOMMER N. & LAPLACE J. (2015). An Avatar Architecture for the Web of Things. *Internet Computing, IEEE*, **19**(2), 30–38.

NIKOLI S., PENCA V. & KONJOVI Z. (2011). Semantic Web Based Architecture for Managing Hardware Heterogeneity in Wireless Sensor Network. In *International Journal of Computer Science and Applications*, volume 8, p. 38–58.

PERERA C., ZASLAVSKY A., LIU C. H., COMPTON M., CHRISTEN P. & GEORGAKOPOULOS D. (2014). Sensor search techniques for sensing as a service architecture for the internet of things. *IEEE Sensors Journal*, **14**(2), 406–420.

POSLAD S., MIDDLETON S. E., CHAVES F., TAO R., NECMIOGLU O. & BUGEL U. (2015). A Semantic IoT Early Warning System for Natural Environment Crisis Management. *IEEE Transactions on Emerging Topics in Computing*, **3**(2), 246–257.

SEYDOUX N., ALAYA M. B., HERNANDEZ N., MONTEIL T. & HAEMMERLÉ O. (2015). Sémantique et Internet des objets : d'un état de l'art à une ontologie modulaire. In *26es Journées francophones d'Ingénierie des Connaissances*.

SHETH A., HENSON C. & SAHOO S. S. (2008). Semantic Sensor Web. In *IEEE Internet Computing*, volume 12, p. 78–83.

VLACHEAS P., GIAFFREDA R., STAVROULAKI V., KELAIDONIS D., FOTEINOS V., POULIOS G., DEMESTICHAS P., SOMOV A., BISWAS A. & MOESSNER K. (2013). Enabling smart cities through a cognitive management framework for the internet of things. *IEEE Communications Magazine*, **51**(6).

ZANELLA A., BUI N., CASTELLANI A., VANGELISTA L. & ZORZI M. (2014). Internet of Things for Smart Cities. *IEEE Internet of Things Journal*, **1**(1), 22–32.

Ontologie

ProVoc : une ontologie pour décrire des produits sur le Web

Cédric Lopez[1], Farhad Nooralahzadeh[2], Elena Cabrio[2], Frédérique Segond[1], Fabien Gandon[2]

[1] Viseo, R&D, Grenoble
{cedric.lopez,frederique.segond}@viseo.com
[2] INRIA, Wimmics, Sophia Antipolis, Nice
{farhad.nooralahzadeh, elena.cabrio, fabien.gandon}@inria.fr

Résumé : De nombreuses recherches ont depuis longtemps motivé l'utilisation d'ontologies pour répondre aux besoins de représentation du e-Commerce. Dans cet article, nous présentons ProVoc (*Product Vocabulary*), une ontologie ayant pour objectif de décrire des produits sur le Web. Complémentaire à GoodRelations (Hepp, 2008), l'ontologie au format du Web sémantique la plus utilisée dans le monde du e-Commerce, Provoc se concentre sur une représentation fine des produits et de leurs entités relatives (gammes des produits, composition des produits, *etc*.). L'utilisation conjointe des deux ontologies permet d'élargir l'espace des requêtes de l'utilisateur. Par exemple : « Quels sont les produits qui contiennent des ingrédients néfastes pour la santé ? Qui les vend ? ». Nous montrons par le biais de requêtes SPARQL que nos scénarios trouvent une formulation adéquate et une représentation pertinente avec ProVoc. Enfin, une application de veille stratégique dans le domaine de la cosmétique est présentée.

Mots-clés : ProVoc, ontologie, web sémantique, représentation des connaissances

1 Introduction

Cette dernière décennie, le nombre de produits disponibles dans le commerce a largement augmenté. Par exemple, le nombre de références pour l'alimentation infantile a augmenté de 58%, le nombre de références pour le café torréfié a augmenté de 81%, quant aux produits de beauté ils ont connu une augmentation de 42%[1].

Devant une telle masse de produits, le client aborde le problème de décision d'achat selon ses propres combinaisons de critères : le prix, la marque, la composition du produit, la qualité, l'appréciation globale par la communauté, les comparatifs, les avis de ses proches, *etc*. Pour aider la prise de décision, de nouvelles applications ont vu le jour, notamment dans le domaine de l'alimentation, telles que ShopWise qui se concentre sur la composition de plus de 25 000 produits alimentaires, EcoCompare qui permet d'évaluer des produits en fonction des critères d'éco-responsabilité (environnement, sociétal, santé), ou SkinDeep qui recense les ingrédients potentiellement dangereux dans les produits cosmétiques.

Alors que les consommateurs cherchent de plus en plus à acquérir des informations sur des produits, les ontologies ouvertes disponibles au format du Web sémantique proposent une représentation pertinente dans le contexte du e-Commerce mais la couverture de la représentation d'informations relatives aux produits eux-mêmes reste faible.

Depuis 2009, Google permet l'enrichissement des résultats de son moteur de recherche en proposant notamment les *rich snippets* dédiés aux e-commerçants. Les *rich snippets* ont pour objectif de fournir un affichage plus pertinent pour les sites Web dans les résultats du moteur

[1] Ces pourcentages sont issus de http://www.journaldunet.com/

(images, évaluations, localisation, *etc.*) et permettent aux développeurs d'améliorer le trafic et le référencement de leurs pages. Les *rich snippet* et les ontologies associées ne couvrent pas nos besoins : le vocabulaire utilisé est un sous-ensemble du vocabulaire de schema.org[2] (précisément des classes *Product*, *Offer*, *AggregateOffer*) lui-même inspiré[3] de GoodRelations (Hepp, 2008) qui concerne essentiellement des scénarios de e-Commerce.

En effet, GoodRelations, l'ontologie au format du Web sémantique la plus utilisée dans le monde du e-Commerce (Ashraf e*t al.*, 2011), se positionne comme le vocabulaire le plus puissant qui permette de publier des détails sur les produits et services. Elle est fondée sur la structure agent-objet-promesse-compensation (agent : personne ou organisation ; promesse : transfert de la propriété d'un objet par exemple ; objet : un objet ou un service ; compensation : par exemple un montant monétaire). Cette ontologie est donc orientée vers les transactions en ligne plus que vers l'aide à la décision d'achat. Adoptant un angle de vue différent, mais compatible, nous proposons l'ontologie ProVoc (*Product Vocabulary*), développée dans le cadre du projet SMILK[4] (*Social Media Intelligence and Linked Knowledge*, LabCom ANR).

Contrairement à GoodRelations qui se place dans le contexte du e-Commerce, Provoc ne porte pas exclusivement sur cette application mais s'intéresse en général à la description et l'organisation de catalogues de produits. De façon complémentaire à GoodRelations, ProVoc présente deux intérêts : 1) une représentation plus fine des produits qui permet de répondre à des requêtes telles que « Quelles gammes contiennent les produits que je recherche ? », 2) la possibilité de tisser des liens vers des informations *a priori* hors catalogue. On peut, par exemple, tisser des liens vers des informations relatives à la santé via la composition des produits (« Quels aliments contiennent des ingrédients néfastes pour la santé ? »), ou tisser des liens vers des ontologies telles que FOAF (Brickley et Miller, 2012) et des bases de connaissances telles que DBpedia. On ouvre ainsi le champs des requêtes possibles : « Quels sont les parfums représentés par des actrices qui ont joué dans Star Wars III ? » et par le biais de GoodRelations « qui les vend et sous quelles modalités ? ». L'objectif de ProVoc est donc de représenter, publier et relier des informations issues de catalogues de produits à d'autres données ouvertes et liées sur le Web ou internes à un Web sémantique d'entreprise.

Dans section suivante, nous adoptons la méthodologie de (Ushold et Gruninger, 1996) pour construire notre ontologie à partir de scénarios issus de clients de la société Viseo. Ces scénarios mettent en avant des situations impossibles à représenter avec GoodRelations que nous résolvons avec ProVoc. La section 3 donne un aperçu[5] des entités et relations de ProVoc et nous discutons leur positionnement vis-à-vis de GoodRelations. Les choix de langages et l'évaluation sont abordés dans la section 4. Un cas d'utilisation concret est présenté dans la section 5 avant de présenter les perspectives de notre projet (section 6).

2 Représentation

Dans la suite, nous identifions les scénarios et questions de compétences en adoptant la méthodologie de (Ushold et Gruninger, 1996). Premièrement, nous identifions des scénarios issus de cas d'utilisation réels identifiés par la société Viseo dans le cadre du laboratoire commun SMILK. A partir de ces scénarios, nous identifions les questions de compétences, c'est-à-dire les questions auxquelles notre ontologie doit être en mesure de répondre.

[2] Schema.org offre un vocabulaire utilisable avec les formats Microdata, RDFa ou JSON-LD pour annoter des informations dans les pages Web.

[3] GoodRelations constitue dès 2012 le cœur de schema.org concernant le e-Commerce.

[4] https://project.inria.fr/smilk/fr/

[5] ProVoc est disponible ici : http://ns.inria.fr/provoc/

2.1 Scénarios

Les scénarios présentés ici sont issus des clients de Viseo du secteur de la cosmétique (L'Oréal, L'Occitane, et Moët Hennessy Louis Vuitton). Certains scénarios sont également issus de notre collaboration avec Beaute-test.com (Lopez *et al.*, 2014), un guide d'achat des cosmétiques en ligne qui fournit près de 50 000 fiches produits et qui a pour objectif d'informer et de conseiller les internautes sur les produits de beauté. De tels scénarios ne peuvent pas être traités par les ontologies existantes.

Ces scénarios motivent et délimitent la conception et la publication du vocabulaire ProVoc et permettent d'identifier des applications associées. Les scénarios 1 à 5 nécessitent un catalogue « précis » des produits ; le scénario 6 montre la nécessité d'établir des liens avec des données « hors catalogue ».

- Scénario 1 : L'utilisateur recherche des informations sur les différents maillons d'une chaîne de distribution d'un produit. L'ontologie doit donc permettre de représenter les entités en mesure de fournir des produits à d'autres entités. Ces entités peuvent être une maison de fabrication, un distributeur[6], *etc.*, toute entité faisant partie intégrante des canaux de distribution.

- Scénario 2 : L'utilisateur cherche à se renseigner sur les composants d'un produit, par exemple, à savoir si ces composants sont néfastes pour la santé. Ces composants peuvent être chimiques, naturels, ou matériels, par exemple. L'ontologie doit donc permettre de représenter la composition des produits et leurs impacts sur la santé.

- Scénario 3 : L'utilisateur souhaite effectuer des recherches par gammes de produits attachées à une marque donnée. Il souhaite notamment pouvoir identifier et naviguer à l'intérieur des gammes. L'ontologie devra permettre de représenter les gammes de produits.

- Scénario 4 : L'utilisateur souhaite connaître les coffrets de produits qui contiennent un ou plusieurs produits en particulier. L'ontologie doit permettre de représenter un ensemble de produits vendus comme une unité. Par exemple un coffret de cosmétique ou un panier alimentaire.

- Scénario 5 : L'utilisateur souhaite connaître la cible d'un produit ou d'une marque (par exemple « animaux », « végétaux », « hommes », « femmes », « adultes » ou « enfants »). L'ontologie devra donc représenter la cible des produits et des marques.

- Scénario 6 : L'utilisateur souhaite acheter un produit représenté par une personne : actrice d'un film donné, mannequin, ou autres personnalités. L'ontologie devra donc représenter les personnes qui sont impliquées dans la publicité des produits.

2.2 Questions de compétences

Dresser une liste de questions de compétences est un moyen de déterminer les spécifications de l'ontologie (Gruninger et Fox, 1995). Une liste non exhaustive des questions de compétence issues des questions posées par les utilisateurs du forum Beauté-test[7] qui a l'avantage d'être très actif et de couvrir de nombreux aspects de la cosmétique (produits de maquillage, produits de soins, parfums, *etc.*), est présentée ici :

- Q1 : Quels sont les fournisseurs d'un produit donné ? (Scénario 1)

[6] (Ding *et al.*, 2004) mettent en avant le fait que les descriptions de produits constituent un élément, et que cet élément doit être le cœur du catalogue, associé à des informations relatives au vendeur, au fabriquant, *etc.*

[7] http://www.beaute-test.com/forums/index.php

- Q2 : Quel(le)s sont les {produits | gammes | marques | divisions | sociétés} qui présentent des risques pour la santé ? (Scénarios 1 et 2)
- Q3 : Quel(le)s sont les {gammes | marques | divisions | sociétés} qui ne commercialisent pas de produits contenant du propylène glycol ? (Scénarios 1 et 2)
- Q4 : Quelles sont les gammes de produits d'une marque donnée ? (Scénario 3)
- Q5 : Quel type de consommateur est ciblé par le/la {produit | gamme | marque} ? (Scénario 5)
- Q6 : Existe-t-il un coffret contenant le produit recherché ? A l'inverse, le produit inclus dans ce coffret est-il commercialisé unitairement ? (Scénario 4)
- Q7 : Quels sont les parfums représentés par des actrices? (Scénario 6)

3 Principales entités et relations de ProVoc

Dans cette section nous présentons et discutons les principales entités et relations de ProVoc. Elles sont issues de la terminologie extraite des questions de compétences, lorsque celles-ci n'étaient pas représentées dans d'autres ontologies[8]. Les primitives centrales du vocabulaire sont présentées en figure 1.

Dans la suite, le préfixe *pv:* désigne des ressources de ProVoc (espace de nommage : http://ns.inria.fr/provoc/), et le préfixe *gr:* désigne des ressources de GoodRelations. L'espace de nommage et la publication respectent les principes des données liées sur le Web et notamment la déréférenciation et la négociation de contenu par HTTP. Le vocabulaire ProVoc est référencé et intégré au catalogue LOV et le préfix « pv » est enregistré sur prefix.cc.

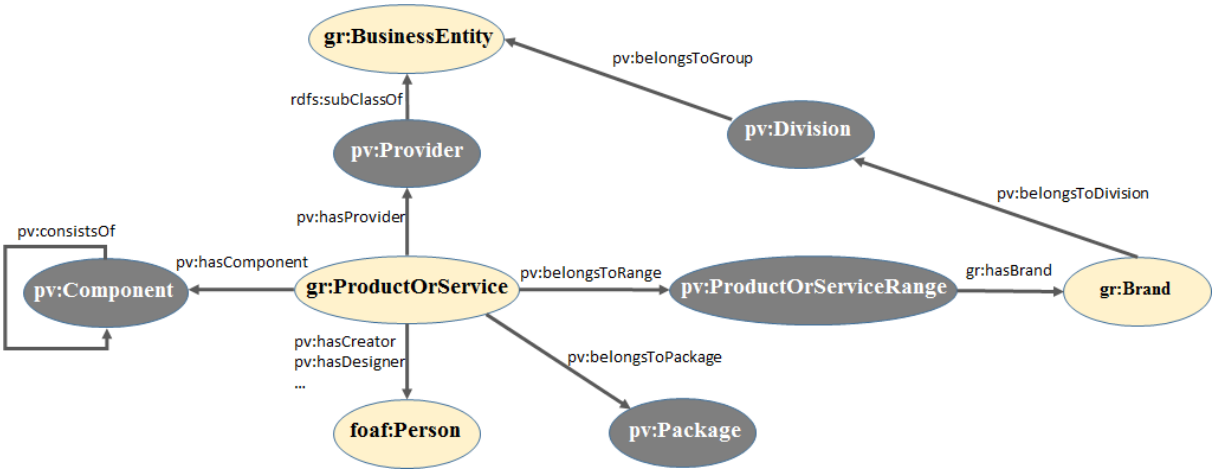


FIGURE 1 – Diagramme représentant les principales classes et propriétés de ProVoc (en gris et préfixées par « pv: »).

[8] La recherche des vocabulaires s'est principalement effectuée via Linked Open Vocabularies (LOV) : http://lov.okfn.org/dataset/lov/.

gr:isVariantOf :
GoodRelations définit des variantes de produits[9] : "*A variant is a specialization of a product model and inherits all of its product properties, unless they are defined locally. This allows a very compact modeling of product models that vary only in a few properties.*"

D'après GoodRelations, une variante d'un MacBook est par exemple un MacBook13Inch ou un MacBook15Inch qui varient par leur taille d'écran et la quantité de ports USB disponibles. Il s'agit d'héritage entre un produit « parent » et ses dérivés qui héritent des caractéristiques par défaut du produit « parent » à moins de redéfinir les valeurs localement, un peu à la manière d'une représentation orientée prototypes. Dans la version actuelle de GoodRelations, il existe une relation *gr:isVariantOf* qui doit nécessairement être utilisée entre deux modèles de produits ou services. Or, les gammes de produits peuvent difficilement être traitées comme un ensemble de dérivés d'un produit/modèle commun. Par exemple, Elsève est une gamme (de la marque L'Oréal Paris) proposant des shampooings avec des dérivés, la même gamme Elsève propose aussi des Colorations avec des dérivés, des Huiles avec des dérivés *etc.* Ainsi, mis à part le trait commun qu'il s'agit de traitements pour les cheveux, ces produits ne partagent pas un prototype commun. L'utilisation de *gr:isVariantOf* entre certains produits proches, impliquerait que l'on obtienne plusieurs ensembles de produits apparentés, au détriment d'une gamme unique.

In fine, les variantes de GoodRelations semblent pertinentes pour identifier des produits plus ou moins similaires[10], mais les gammes de produits ont d'après nous une toute autre vocation, notamment d'un point de vue fonctionnel et marketing, impliquant qu'elles doivent être définies par le fournisseur de façon non subjective. Or, *gr:isVariantOf* a une sémantique très large et subjective[11]. Par exemple, rien n'empêche d'exprimer qu'une Renault Clio 4 est une variante d'une Ford Fiesta[12] ; pourtant elles ne sont pas de la même marque.

Pour ces raisons, nous introduisons dans ProVoc la notion de gammes de produits ou services *pv:ProductOrServiceRange*.

pv:ProductOrServiceRange *:*
pv:ProductOrServiceRange permet de représenter une gamme de produits de façon non subjective (fournie par l'expert) contrairement à *gr:isVariantOf*. Une *pv:ProductOrServiceRange* ne représente ni un produit ni un modèle de produit. L'utilisation de *pv:ProductOrServiceRange* permet d'affiner la représentation et le contenu des catalogues de produits et services tout en leur assurant un caractère objectif. On peut ainsi exprimer qu'un produit appartient à une gamme et que cette gamme est proposée par une marque. Un exemple d'utilisation est donné Figure 2.

```
@prefix pv: <http://ns.inria.fr/provoc#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix ex: <http://example.org#> .

ex:Huile Extraordinaire Universelle
              pv:belongsToProductOrServiceRange   ex:Elsève .
ex:Elsève     rdf:type   pv:ProductOrServiceRange ;
              pv:belongsToBrand   ex:LOreal_Paris .
```

FIGURE 2 – Exemple d'utilisation (en Turtle) de la classe pv:belongsToProductOrServiceRange pour le cas de la gamme Elsève.

[9] http://wiki.goodrelations-vocabulary.org/Documentation/Product_variants, http://purl.org/goodrelations/v1#isVariantOf

[10] voir également http://purl.org/goodrelations/v1#isSimilarTo qui ne se différencie de *gr:isVariantOf* que par le fait que le domain et le range de *gr:isSimilarTo* sont plus large

[11] C'est également le cas de *gr:isSimilarTo*, tel que le commente M. Hepp : http://purl.org/goodrelations/v1#isSimilarTo

[12] http://www.autonews.fr/nouveautes/nouveaute/104216-renault-clio-fiesta-prix/

pv:Component :
Une instance de cette classe représente un composant d'un produit. Un *pv:Component* peut être constitué d'autres *pv:Component*. Par exemple un volant ou un tuyau d'échappement pour une voiture, les ingrédients d'un parfum, *etc.*

pv:Division :
Une instance de cette classe représente une division (un sous-groupe) de BusinessEntity. En effet, une organisation est parfois divisée en plusieurs divisions, et chaque division propose des marques différentes. GoodRelations lie *gr:BusinessEntity* directement à *gr:Brand*. Par exemple, L'Oréal Grand Public est une division du groupe L'Oréal.

pv:Package :
Un package est un ensemble de produits et/ou services. Par exemple, un coffret de cosmétique qui contient des crèmes, un parfum, et rouge-à-lèvres. Un autre exemple associant un produit et un service pourrait être un package contenant un téléphone et son abonnement.
Cette classe est utilisée pour représenter un ensemble de produits qui est vendu unitairement. Cet ensemble représenté par *pv:Package* peut contenir des variantes d'un modèle de produit (utilisation de *gr:isVariantOf*), des produits appartenant à une même gamme (utilisation de *pv:ProductOrServiceRange*), ou des produits similaires (utilisation de *gr:isSimilarTo*). Un exemple est donné Figure 3.

```
@prefix pv: <http://ns.inria.fr/provoc#> .
@prefix gr: <http://purl.org/goodrelations/v1#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix ex: <http://example.org#> .

ex:Degustabox1015   rdf:type   pv:Package .
ex:MiniBiscuitsBanana  pv:belongsToPackage   ex:Dugustabox1015 ;
                       pv:belongsToBrand     ex:Weetabix Crispy Minis ;
                       gr:isSimilarTo        ex:MiniBiscuitsChocolate ;
                       gr:isVariantOf        ex:MiniBiscuitsModel .
ex:MiniBiscuitsModel   rdf:type              gr:ProductOrServiceModel .
ex:CappucinoCarambar   pv:belongsToPackage   ex:Dugustabox1015 ;
                       pv:belongsToBrand     ex:Maxwell House .
ex:ExtraitNaturelDeVanille   pv:belongsToPackage   ex:Dugustabox1015 ;
                             pv:belongsToBrand     ex:Sainte-Lucie .
```

FIGURE 3 – Exemple d'utilisation (en Turtle) des classes *pv:Package*, *gr:isVariantOf* et *gr:isSimilarTo* dans le cas d'un panier alimentaire[13] contenant 3 produits.

pv:Provider : Le fournisseur est un type d'organisation. Il se distingue de l'organisation par le fait que les marques ne lui appartiennent pas ; il ne fait que les commercialiser. Le fournisseur propose des marques à la vente qui n'appartiennent pas toujours à la même organisation. Exemple : Carrefour vend des produits des entreprises Lustucru et Sony.

D'autres ressources viennent enrichir ce modèle, notamment pour assurer les liaisons entre les données du catalogue et les données hors-catalogue. Par exemple, *pv:Ambassador*, *pv:Designer*, *pv:Model*, sont des sous-classes de *foaf:Person* qui représentent des fonctions qu'exercent des personnes impliquées dans la chaîne de production, la commercialisation, ou la communication du produit (par exemple les égéries des marques). L'introduction de ces classes dans ProVoc permet d'établir les liens entre le catalogue de produit et des données hors

[13] extrait de représentation d'un panier alimentaire : http://laboxdumois.fr/box/degustabox-octobre-2015.html

catalogue (ici en se référant à des personnes), provenant de bases de connaissances telles que DBpedia.

Le tableau 1 recense les 19 propriétés de ProVoc.

TABLE 1 – *Liste non exhaustive des propriétés de Provoc*

Propriétés	**Domaine**	**Co-domaine**
pv:belongsToBrand	pv:ProductOrServiceRange, gr:ProductOrService	gr:Brand
pv:belongsToDivision	gr:Brand	pv:division
pv:belongsToGroup	pv:Division	gr:BusinessEntity
pv:belongsToPackage	gr:ProductOrService	pv:Package
pv:belongsToProductOrServiceRange	gr:ProductOrService	gr:ProductRange
pv:consistsOf	pv:Component	pv:Component
pv:hasComponent	gr:ProductOrService	pv:Component
pv:hasCreator	gr:ProductOrService	foaf:Person
pv:hasFragranceCreator	gr:ProductOrService	foaf:Person
pv:hasPackageDesigner	gr:ProductOrService	foaf:Person
pv:hasFounder	gr:BusinessEntity	foaf:Person
pv:hasRepresentative	Ø	foaf:Person
pv:hasProvider	gr:ProductOrService	gr:Provider
pv:hasAmbassador	gr:ProductOrService	foaf:Person
pv:hasModel	Ø	foaf:Person
pv:hasFunctionality	pv:Component	string
pv:hasTarget	gr:ProductOrService	Ø
pv:hasVersion	gr:ProductOrService	string
pv:healthImpact	pv:Component	Ø

4 Choix du langage et évaluation

ProVoc a été éditée avec le logiciel Protégé[14]. L'ontologie et la description de ses ressources et propriétés sont publiées selon les principes des données liées sur le Web et le schéma est identifié par l'URI HTTP http://ns.inria.fr/provoc#.

L'ontologie ProVoc utilise les mêmes primitives que GoodRelations : owl:Ontology, owl:Class, owl:versionInfo, owl:DatatypeProperty, rdfs:subClassOf, rdfs:subPropertyOf, rdfs:comment, rdfs:domain, rdfs:range, rdf:datatype, rdf:type. De cette façon, les annotations effectuées avec ProVoc et GoodRelations sont dans le même fragment d'expressivité et peuvent être interprétées par un raisonneur RDF(S) qui sait traiter les éléments mentionnés ci-dessus.

Nous montrons, par le biais de quelques exemples de requêtes SPARQL, que toutes nos questions de compétences trouvent une réponse. Dans les requêtes suivantes, on notera le préfixe *pv: <http://ns.inria.fr/provoc#>*.

Pour la question de compétence Q2 « Quel(le)s sont les {produits | gammes | marques | divisions | sociétés} qui présentent des risques pour la santé ? » issue des scénarios 1 et 2, des exemples de questions concrètes sont :

[14] http://protege.stanford.edu/

- Exemple 1 : « Quels sont les types de produits qui contiennent du linalool ? »

  Sa formulation en SPARQL est :

```
SELECT DISTINCT ?type
WHERE {
      ?product  pv:contains <http://fr.dbpedia.org/page/Linalool> .
      ?product rdf:type ?type
}
```

- Exemple 2 : « Quelles sont les gammes de produits de la marque L'Oréal Paris qui présentent des risques pour la santé ? »

  Sa formulation en SPARQL est :

```
SELECT DISTINCT ?range
WHERE {
      ?range pv:belongsToBrand ex:LOreal Paris .
      ?product pv:belongsToRange ?range .
      ?product pv:contains ?component .
      ?component pv:healthImpact ?healthImpact
}
```

Pour la question de compétence Q4 « Quelles sont les gammes de produits d'une marque ? » issue du scénario 3, un exemple de question concrète est :

- Exemple 4 : Quelles sont les gammes de produits proposées par L'Oréal Paris ?

  Sa formulation en SPARQL est :

```
SELECT DISTINCT ?range
WHERE {
      ?range pv:belongsToBrand ex:LOreal Paris
}
```

Pour la question de compétence Q5 « Quel type de consommateur est ciblé par le/la {produit | gamme | marque} ? » issue du scénario 5, un exemple de question concrète est :

- Exemple 5 : Quel type de consommateur est ciblé par la gamme Elsève ?

  Sa formulation en SPARQL est :

```
SELECT DISTINCT ?consumer
WHERE {
      ex:Elseve pv:hasTarget ?consumer
}
```

Pour la question de compétence Q6 « Existe-t-il un coffret contenant le produit recherché ? A l'inverse, le produit inclus dans ce coffret est-il commercialisé unitairement ? » issue du scénario 4, un exemple de question concrète est :

- Exemple 6 : Quels sont les coffrets distribués par Sephora qui contiennent le parfum La vie est Belle ?

Sa formulation en SPARQL est :

```
SELECT DISTINCT ?package
WHERE {
      ex:La vie est belle pv:belongsToPackage ?package .
      ?package pv:hasProvider ex:Sephora
}
```

Pour la question de compétence Q7 « Quels sont les parfums représentés par des actrices ? » issue du scénario 6, un exemple de question concrète est :

- Exemple 7 : Quels sont les produits représentés par des acteurs qui ont joué dans Star Wars III ?

Sa formulation en SPARQL est (avec le préfixe *dbo:http://dbpedia.org/ontology/*) :

```
SELECT DISTINCT ?product
WHERE {
      <http://dbpedia.org/page/Star_Wars_Episode_III:_Revenge_of_the_Sith>
dbo:starring ?actor .
      ?product pv:hasRepresentative ?actor .
}
```

En intégrant le vocabulaire de GoodRelations, ces requêtes peuvent être enrichies par des interrogations relatives au e-Commerce. Par exemple, en cosmétique, « Qui vend, et sous quelles conditions, des coffrets destinés aux hommes contenant des produits appartenant à des gammes qui n'utilisent aucun composant néfaste pour la santé ? ».

5 Cas d'application

Chez Viseo, le secteur de la cosmétique est bien représenté par des clients d'envergure tels que L'Oréal, L'Occitane, ou LVMH (Moët Hennessy Louis Vuitton). Dans un contexte de veille stratégique, nous avons développé un prototype prenant la forme d'un plugin de navigateur ayant une double ambition :

- Enrichir les connaissances des utilisateurs naviguant sur le Web à l'aide de résultats issus du Traitement Automatique du Langage Naturel (TAL), du Web de Données et des réseaux sociaux.
- Peupler une base de connaissance utilisant, entre autres, le vocabulaire ProVoc permettant de répondre aux questions issues des scénarios définis à la section 2.

Le prototype (cf. Fig. 4) analyse les pages Web en vue d'identifier les entités nommées du domaine de la cosmétique, par exemple les noms de produits, les noms de gammes de produits ou les groupes de cosmétique ; celles-ci correspondent aux classes de ProVoc. La reconnaissance automatique de telles entités et de leurs relations est effectuée par Renco, un système à base de règles linguistiques (Lopez *et al.*, 2014). Le système peuple ainsi semi automatiquement la base de connaissance ProVoc qui peut être visionnée (avec ses liens vers

DBpedia et NetScent[15]) et interrogée dans notre outil Viewer SMILK qui repose sur le serveur RDF Jena Fuseki.
La totalité des requêtes issues des questions de compétences exprimées dans la section 2 a été exécutée avec succès.

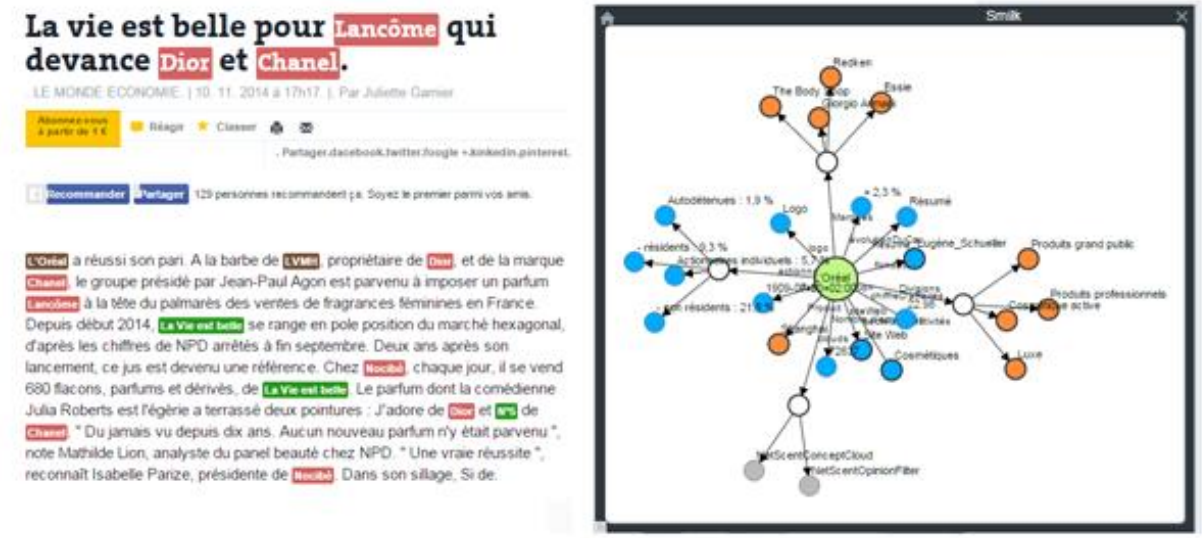


FIGURE 4 – Vue du prototype SMILK qui analyse une page de www.lemonde.fr : à gauche, reconnaissance d'entités d'intérêts ; à droite, exemple de graphe RDF pour « L'Oréal », généré à partir des bases de connaissance ProVoc (en bleu), DBpedia (en orange), NetScent (en gris).

6 Conclusion

Alors que de nombreux résultats de recherche ont depuis longtemps motivé l'utilisation d'ontologies pour le e-Commerce (Glushko *et al.*, 1999 ; Fensel *et al.*, 2001), il a fallu attendre 2008 pour voir apparaître GoodRelations, la première ontologie respectant les standards du Web Sémantique pour répondre aux besoins de représentation du e-Commerce. Dans cet article, nous avons présenté notre travail qui a consisté à modéliser, représenter et publier l'ontologie ProVoc, une extension de GoodRelations.
Nous avons vu comment ProVoc est utilisable pour représenter des catalogues de produits, en considérant des classes telles que les gammes de produits, des paquets de produits vendus en tant qu'unité, ou encore les composants de produits, entre autres. ProVoc n'a pas pour ambition de représenter une taxonomie des produits existants. Ainsi, les produits représentés par ProVoc peuvent être typés (*rdf:type*) en utilisant d'autres ontologies telles qu'eClassOWL ou unspscOWL (Hepp, 2005 ; Hepp, 2007) en fonction du cas d'application envisagé.
Associée à GoodRelations, ProVoc apporte aux sociétés une meilleure visibilité de leurs produits, plus fine, ainsi qu'une plus grande transparence sur les produits commercialisés qui offre à l'utilisateur/consommateur de nouveaux modes de recherches et la possibilité d'exploiter plus de critères pour exprimer des requêtes répondant à ses attentes.
Comme le mentionnent (Ding *et al.*, 2004), il est improbable qu'un standard puisse couvrir tous les aspects du e-Commerce pour tous les marchés verticaux, ce qui conduira très certainement à l'enrichissement de l'ensemble des vocabulaires déjà proposés.

[15] Base d'opinions dans le contexte de la cosmétique, développée par Holmes Semantic Solutions (http://www.ho2s.com/)

7 Remerciements

Ce travail est réalisé dans le cadre du Laboratoire Commun SMILK financé par l'ANR (ANR-13-LAB2-0001).

Références

ASHRAF J., CYGANIAK R., O'RIAIN S., & HADZIC M. (2011). Open eBusiness Ontology Usage: Investigating Community Implementation of GoodRelations. Workshop on *Linked Data On the Web*.

BRICKLEY D., & MILLER L. (2012). FOAF vocabulary specification 0.98. *Namespace document*, *9*.

DING Y., FENSEL D., KLEIN M., OMELAYENKO B., & SCHULTEN, E. (2004). The role of ontologies in ecommerce. In *Handbook on ontologies*, p. 593-615. Springer Berlin Heidelberg.

FENSEL D., MCGUINESS D. L., NG W. K., & YAN G. (2001). Ontologies and electronic commerce. *Intelligent Systems, IEEE*, 16(1), p. 8-14.

GLUSHKO R. J., TENENBAUM J. M., & MELTZER B. (1999). An XML framework for agent-based E-commerce. *Communications of the ACM*, 42(3), p. 106-114.

GRUNINGER M., & FOX M. S. (1995). Methodology for the Design and Evaluation of Ontologies. In: *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95*, Montreal.

HEPP M. (2005). eClassOWL: A fully-fledged products and services ontology in OWL. *Poster Proceedings of ISWC2005. Galway.*

HEPP M., & DE BRUIJN, J. (2007). GenTax: A generic methodology for deriving OWL and RDF-S ontologies from hierarchical classifications, thesauri, and inconsistent taxonomies. In *The semantic web: research and applications*, p. 129-144.

HEPP M. (2008). Goodrelations: An ontology for describing products and services offers on the web. In *Knowledge Engineering: Practice and Patterns*, p. 329-346. Springer Berlin Heidelberg.

LOPEZ C., SEGOND F., HONDERMARCK O., CURTONI P., & DINI, L. (2014) Generating a Resource for Products and Brandnames Recognition. Application to the Cosmetic Domain. *Proceedings of LREC2014*, p. 2559-2564.

USCHOLD M., & GRUNINGER M. (1996). Ontologies: Principles, methods and applications. *The knowledge engineering review*, 11(02), p. 93-136.

Utilisation d'ontologies pour la quête de vérité : une étude expérimentale

Valentina Beretta[1], Sébastien Harispe[1], Sylvie Ranwez[1], Isabelle Mougenot[2]

[1] LGI2P de l'école des mines d'Alès, Site de Nîmes, Parc G. Besse, F-30 035 Nîmes
{prenom.nom}@mines-ales.fr
[2] UMR Espace-Dev, Université de Montpellier, Rue JF. Breton, Montpellier, France
isabelle.mougenot@umontpellier.fr

Résumé : L'objectif principal des méthodes de recherche de vérité (*truth-finding* en anglais) consiste à déterminer les valeurs les plus fiables et dignes de confiance parmi celles qui sont associées à un ensemble de faits. La plupart des méthodes actuelles supposent qu'il n'existe qu'une seule valeur 'vraie', ce qui les rend inappropriées pour des applications réelles où plusieurs valeurs peuvent être considérées comme vraies simultanément : Paris est une capitale, mais également une ville ou même un département…
Cet article propose une extension de la formalisation habituellement utilisée dans la littérature, afin de prendre en compte les relations entre les valeurs candidates définies au sein d'une ontologie de domaine et ainsi déterminer les valeurs qui entrent en conflit et celles qui peuvent être 'vraies' simultanément. Notre approche s'inspire des fonctions de croyance afin de propager une valeur de confiance à ces valeurs candidates en fonction de l'ordre partiel établi par l'ontologie. L'évaluation de notre approche, tenant compte de faits extraits de DBpédia, démontre son efficacité par rapport aux approches classiques – au travers notamment d'une diminution du taux d'erreur pouvant aller de 16 à 30%.

Mots-clés : Détection de vérité, ontologie, confiance, fiabilité des sources, fonctions de croyance.

1 Introduction

Aucune source d'information n'est aussi prolifique que le Web, et ce pour plusieurs raisons. D'une part la collecte et le partage de données sont facilités par des avancées technologiques (e.g. objets connectés). D'autre part la liberté de publication incite chacun à être *fournisseur de contenu*, par exemple sur les réseaux sociaux ou les plateformes collaboratives gratuites et publiques telles que Wikipédia, pour ne citer que quelques exemples. Or ce qui parait être un avantage pour de nombreuses applications qui tirent parti des ressources accessibles sur le Web pour peupler des bases de connaissance, inférer de la connaissance ou mener une analyse des habitudes commerciales de certains clients, par exemple, peut vite atteindre ses limites si la validité des informations n'est pas prise en compte. Dans ce cas, les moteurs d'inférence et les raisonneurs peuvent amener à de mauvaises conclusions et impacter négativement les performances de certains systèmes, voire inciter l'utilisateur à prendre de mauvaises décisions. C'est ce qui a conduit à l'émergence de nombreux travaux dédiés à la *recherche de vérité* (ou *détection de vérité – truth-finding* en anglais). L'objectif est relativement simple : trouver les données qui semblent être probables et, de façon intimement liée, distinguer les sources d'information les plus *fiables*. En effet, l'un des meilleurs indicateurs de la *confiance* à associer à une donnée reste sa provenance. L'estimation de la fiabilité d'une (source d')information est du plus grand intérêt pour des domaines tels que la recherche d'information, la détection d'opinion ou encore l'aide à la décision.

Evaluer la fiabilité d'une source est une tâche complexe car elle dépend de nombreux facteurs qui peuvent difficilement être intégrés dans un modèle unique (Fogg & Tseng, 1999; Gil & Artz, 2007; Kelton *et al.*, 2008). Certains de ces facteurs sont directement liés à la source elle-même : crédibilité, couverture, spécificité, popularité, validité et stabilité. D'autres dépendent de l'entité ou des personnes en charge d'affecter un degré de confiance à une source : degré d'expertise, exploitation attendue du degré de fiabilité...

Du fait de l'importance de la tâche de découverte de vérité, la littérature recense de nombreux travaux qui y sont dédiés. On peut y distinguer deux principales approches : les modèles basés sur la *renommée* d'une source et ceux qui utilisent des techniques de recherche de vérité sur les données elles-mêmes. Les méthodes basées sur la renommée s'intéressent à des signaux extérieurs à la source, comme par exemple les liens entre les sources, les logs, les statistiques sur les clics ou encore l'analyse de *spam*. Les autres approches s'intéressent au contenu même de l'information. C'est dans ce deuxième contexte que s'inscrivent nos travaux.

L'objectif principal de la *découverte de vérité* est d'identifier la 'vérité', parmi un ensemble de propositions (*faits*[1]) potentiellement contradictoires. Le principe de base est de considérer que les sources qui fournissent des informations vraies le plus souvent, vont être associées à un fort degré de fiabilité et que les informations fournies par de telles sources sont supposées dignes de confiance (Y. Li et al., 2015). Ainsi, un processus itératif peut être mis en place pour déterminer ces différents degrés de fiabilité.

La plupart des modèles existants partent du postulat qu'une seule valeur peut être vraie parmi celles proposées par les différentes sources. Pourtant, dans la plupart des cas, les valeurs proposées ne sont pas indépendantes. Un ordre partiel sur ces valeurs peut exister. Par exemple parmi les propositions suivantes, deux valeurs seulement entrent en conflit :

- *<Pablo Picasso, bornIn, Spain>*
- *<Pablo Picasso, bornIn, Europe>*
- *<Pablo Picasso, bornIn, Malaga>*
- *<Pablo Picasso, bornIn, Granada>*

En effet, Granada et Malaga étant deux villes distinctes, elles ne peuvent être considérées toutes les deux comme étant vraies. Or avec une connaissance ontologique du domaine, il est possible de déterminer que Malaga et Granada sont toutes les deux en Espagne et donc en Europe. C'est cette connaissance que nous désirons ajouter aux modèles existants afin de prendre en compte ce type de relations.

Nos contributions sont les suivantes : i) proposer une nouvelle formalisation du problème de la détection de vérité qui prenne en compte la connaissance du domaine, ii) décrire les adaptations des modèles existants qui sont nécessaires pour intégrer cette connaissance, iii) proposer une évaluation de l'adaptation d'une approche existante.

Après avoir présenté l'état de l'art et notre positionnement dans la section suivante, la section 3 détaillera la formalisation du problème et l'approche proposée. La section 4 présente nos résultats en deux temps : tout d'abord la génération d'un jeu de tests adapté au nouveau contexte de la détection de vérité que nous proposons, puis les expérimentations que nous avons menées basées sur ce jeu de tests et les résultats obtenus. La section 5, enfin, synthétise notre approche et ouvre de nombreuses perspectives de recherche.

2 Positionnement et état de l'art

Par souci de clarté, cette section définit les notations utilisées par la suite. Certaines sont couramment utilisées dans le domaine (Y. Li et al., 2015; Waguih & Berti-Equille, 2014; Yu, 2008), alors que les autres sont introduites pour être utilisées ensuite dans la description de notre approche.

[1] On appelle *fait* un triplé <objet, prédicat, valeur>.

Soit $o \in O$ un *objet* d'intérêt, par exemple 'Pablo Picasso' et $d \in D$ une *description*[2] de cet objet, i.e. un *prédicat* associé à cet objet, par exemple 'Pablo Picasso – bornIn'. Nous appelons *fait* toute paire $f \in F$ de la forme (d, v) où $v \in V$ représente la valeur associée à la description d. Les faits sont émis par des sources qui peuvent être des sites Web, des publications, des articles de journaux, etc. L'ensemble de ces sources est noté S. La découverte de vérité consiste alors à résoudre les conflits qui peuvent exister entre différents faits émis par des sources distinctes. Ainsi, chaque description d peut être associée à un ensemble de faits $F_d \subseteq F$ exprimé au travers de différentes sources $S_d \subseteq S$ ce qui permet de définir l'ensemble $V_d \subseteq V$ qui représente toutes les valeurs qui peuvent être associées à d. Chaque source $s \in S$ exprime un certain nombre de faits $F^s \subseteq F$ et un même fait peut être proposé par plusieurs sources $S^f \subseteq S$.

Pour résoudre les conflits potentiels entre différents faits, il est nécessaire de prendre en compte la fiabilité des sources. On utilise pour ce faire deux fonctions : la *fiabilité* d'une source, que nous noterons t[3], et la *confiance* dans un fait que nous noterons c. Ces fonctions sont définies comme suit.

- $t: S \rightarrow [0,1]$, la *fiabilité* d'une source, représente sa propension à fournir des valeurs vraies. Dans la littérature c'est parfois le terme *poids* qui est utilisé (Y. Li et al., 2015). Une source réputée sûre aura un fort degré de fiabilité et sera considérée comme exprimant des valeurs vraies ($t(s) \simeq 1$) alors qu'une source non sûre aura un degré de fiabilité faible ($t(s) \simeq 0$) et sera réputée pour exprimer des valeurs fausses.
- $c: F \rightarrow [0,1]$, la *confiance* dans un fait, traduit sa propension à être correct, en fonction de nos connaissances actuelles (contexte). En effet, la vérité absolue n'existe pas et ce qu'on qualifie de *vrai*, ne l'est souvent qu'à la lumière de nos connaissances du monde (Pasternack & Roth, 2010). Un fait exact va avoir un fort degré de confiance ($c(f) \simeq 1$) et sera supposé provenir d'une source fiable. Par ailleurs, un fait inexact aura un faible degré de confiance ($c(f) \simeq 0$) et sera supposé provenir d'une source peu fiable.

On voit dans ces deux définitions, l'étroite relation qui existe entre fiabilité et confiance.

A l'aide de ces notations, il est possible de définir la découverte de vérité comme suit – cette définition est une adaptation de celle qui est donnée dans (Y. Li et al., 2015) afin de conserver la cohérence de notation dans la suite de l'article.

Définition 1 – Soit un ensemble de descriptions D, un ensemble de valeurs V, un ensemble de sources S et un ensemble de faits $F \subseteq D \times V$, composé de tous les faits proclamés par toutes les sources de S pour chaque description de D, i.e. $F = \bigcup_{s \in S, d \in D} F_d^s$. Une valeur spécifique fournie par une source s à propos d'une description d est représentée par $v_d^s \in V_d$. L'objectif principal de la découverte de vérité est de trouver pour tous les $d \in D$, $v_d^* \in V_d$, la valeur vraie parmi un ensemble de valeurs associées à cette description[4]. Dans le même temps, les méthodes de détection de vérité estiment la fiabilité des sources, $t(s)$, qui pourra influencer la détection de vérité.

Les différentes approches proposées dans la littérature pour l'identification de vérité peuvent être classées en trois catégories que nous nommons les approches de *référence*, les approches *basiques* et les approches *étendues*.

Les *approches de référence* utilisent des règles de vote entre les différentes sources (Y. Li et al., 2015). Ces approches font l'hypothèse que chaque source a le même degré de fiabilité. Ainsi, la valeur considérée comme vraie sera celle qui apparait le plus grand nombre de fois dans les différentes sources. Ce modèle, très simple, possède deux limites majeures : chaque

[2] Nous employons le terme *description* comme traduction de *data item* couramment utilisé dans la littérature anglaise.

[3] En effet en anglais nous parlerons de *trustworthiness* et cela évite la confusion avec la confiance associée à des faits.

[4] Par exemple, on peut écrire $d = (Pablo\ Picasso, bornIn)$, $v_d^* = Spain$, $f = \big((Pablo\ Picasso, bornIn), Spain\big)$.

source est considérée de la même façon, même celles qui pourraient être qualifiées de non-fiables sur le long terme, et ces approches sont très sensibles à des attaques de type *spam*.

Les *approches basiques* prennent en compte la fiabilité des sources. Pour cela, elles procèdent suivant le modèle itératif présenté dans la section précédente. La confiance dans un fait est estimée en prenant en compte la fiabilité des sources et pour chaque source, sa fiabilité est mise à jour en fonction de la véracité des faits qui lui sont associés. Les principales approches de cette catégorie sont : *Sums*, *AverageLog*, *Investment* et *PooledInvestment* décrites dans (Pasternack & Roth, 2010), et *Cosine* et *2-Estimated* décrites dans (Galland *et al.*, 2010). Elles se distinguent par les formulations employées et la procédure itérative utilisée. De plus chaque approche relaxe certaines hypothèses et se concentre sur des aspects particuliers. Par exemple certaines approches prennent l'hypothèse d'une totale indépendance entre les faits (Y. Li et al., 2015), alors que d'autres utilisent des méthodes de vote complémentaires (Galland *et al.*, 2010). Aucune de ces approches ne considère la connaissance du domaine dans leur processus de détection.

Des *approches étendues* ont donc été proposées, qui prennent en compte des possibles dépendances entre les faits exprimés. La plupart de ces approches analysent des dépendances statiques (Blanco *et al.*, 2010; Dong *et al.*, 2010; Dong *et al.*, 2009a; Pochampally *et al.*, 2014; Qi *et al.*, 2013; Wang *et al.*, 2015) et une approche est proposée pour prendre en compte la dépendance temporelle (Dong *et al.*, 2009b). Dans cette dernière, les changements de dépendance au cours du temps sont considérés (suivi des mises à jour). Dans toutes ces méthodes, l'intuition qui est suivie est que les sources qui partagent les mêmes valeurs fausses sont supposées être interdépendantes. Par exemple, la recopie d'une source sur une autre est estimée (nombreuses redites entre un site et un autre, par exemple). Cette ressemblance entre les sources peut s'observer au niveau des sources elles-mêmes ou d'un groupe de sources. D'autres modèles étendus intègrent une connaissance complémentaire : des similarités entre valeurs, une connaissance antérieure, des techniques de raisonnements, ou encore de l'extraction d'information. *TruthFinder*, par exemple, ajuste son calcul de confiance en un fait, en utilisant une similarité (Yu, 2008). Cette similarité est estimée entre des valeurs numériques, ou des chaînes de caractères, par exemple. Dans (Zhao *et al.*, 2012) la distribution des qualités des sources est prise en compte. *3-Estimates* introduit la notion de *solidité* des faits, c'est-à-dire intégrer dans le calcul de fiabilité d'une source la propension d'un fait à être associé à une valeur fausse (Galland *et al.*, 2010). Dans (Pasternack & Roth, 2011) d'autres informations complémentaires sont prises en compte. Par exemple l'exactitude des extracteurs, la similarité entre faits ou encore l'appartenance à certains groupes de faits. Cette dernière est également utilisée dans (Gupta *et al.*, 2011). L'idée principale consiste à considérer la fiabilité des sources uniquement pour les objets appartenant à un sous-ensemble de sources considérées fiables. Enfin, dans (Dong *et al.*, 2015) l'erreur commise par les extracteurs automatiques est prise en compte.

A notre connaissance, très peu d'approches s'intéressent à des prédicats non-fonctionnels, ceux pour lesquels plusieurs valeurs peuvent être possibles simultanément pour une description donnée, par exemple quand plusieurs personnes sont auteur d'un même livre (Pochampally *et al.*, 2014; Wang *et al.*, 2015; Zhao *et al.*, 2012). Ces approches considérant de multiples vérités sont évaluées par des mesures de *précision* et de *rappel* et partent du postulat qu'une source peut émettre plus d'un fait pour chaque aspect du monde réel (chaque description). Les modèles existants ne considèrent pas la connaissance antérieure que l'on peut avoir sur certaines valeurs. Il est à noter que ces approches sont complètement différentes de celle qui est proposée dans la section suivante. En effet, nous considérons des prédicats fonctionnels, i.e. pour lesquels il n'y a qu'une seule valeur 'vraie', mais pour laquelle la structuration de la connaissance permet de définir un ensemble de valeurs 'vraies' représentant des granularités différentes, des points de vue différents sur cette valeur.

3 Formalisation du problème et description de l'approche proposée

Dans un premier temps nous allons reformuler la problématique de façon à ce qu'elle prenne en compte la définition d'une connaissance du domaine ; puis nous détaillerons l'approche que nous avons adoptée pour rechercher la vérité parmi un ensemble de faits.

3.1 Reformulation de la problématique

Rappelons que nous considérons ici l'analyse de faits associés à des prédicats fonctionnels. Afin de sélectionner la valeur vraie associée à une description, tout comme pour estimer la confiance associée à une source, nous considérons que les valeurs proposées par les sources respectent la logique bivalente, et sont donc *vraies* ou *fausses*. La notion de vérité peut donc être définie par la fonction binaire suivante :

$$tf : F \rightarrow \{true, false\} \quad (1)$$

La formulation du problème que nous proposons vise à représenter de façon plus réaliste les cas réels pour lesquels la dépendance entre plusieurs valeurs est prise en compte. Comme nous allons le voir, cette considération implique des modifications importantes dans la formulation du problème ; cela, aussi bien au niveau des assomptions considérées qu'au niveau des solutions proposées pour résoudre le problème. Nous considérons que la dépendance entre les différentes valeurs est précisée *a priori* dans une ontologie[5], sous la forme d'un ordre partiel $O = (\preccurlyeq, V)$ structurant les différentes valeurs au travers de relations transitives. L'ordre partiel O précise les relations de l'ontologie qui sont prises en compte entre les valeurs, i.e. les relations qui permettent de préciser les valeurs qui subsument d'autres valeurs. Ainsi, pour les valeurs $x, y \in V^2$, écrire $y \preccurlyeq x$ signifie que y implique x. Par exemple $Espagne \preccurlyeq Europe$ signifie que dire que quelqu'un est né en $Espagne$ implique de dire que cette personne est née en $Europe$. Ici nous considérons uniquement l'ordre partiel défini par les relations transitives. Nous ne discuterons pas les notions supplémentaires relatives à la sémantique associée aux relations pouvant exister entre les différentes valeurs. L'ordre partiel peut ainsi être une taxonomie composée de triplets contenant la relation **`subClassOf`**, la relation **`partOf`**, ou représenter un graphe orienté acyclique associé à une sémantique plus complexe. Dans tous les cas, cet ordre partiel pourra être intégré à l'analyse des faits exprimés par les sources étudiées, comme connaissances supplémentaires sur les valeurs considérées. En effet, si une source exprime un fait, elle supporte aussi de façon implicite l'ensemble des faits qui le subsume. Plus formellement, une source exprimant un fait $f = (d \in D, x \in V)$ supporte aussi l'ensemble des faits f' associés à la description d qui impliquent des valeurs plus générales que x, i.e. $\forall f' \in d \times \{y | x \preccurlyeq y\}$, $f \Rightarrow f'$. En effet, un fait étant défini comme une paire $(d \in D, x \in V)$, quand d est connu, un ordre partiel sur les faits peut être appliqué à partir de l'ordre partiel défini sur les valeurs. Dans la suite, par abus de langage, nous utiliserons indifféremment *fait* ou *valeur* quand la description d est connue et fixe.

Si l'on se place dans ce contexte, la valeur de vérité ne peut être réduite à une valeur unique mais se compose plutôt d'un ensemble de valeurs. Si l'on reprend l'exemple décrit en Section 1, les deux faits <*Pablo Picasso, bornIn, Granada*> et <*Pablo Picasso, bornIn, Malaga*> supportent les deux faits <*Pablo Picasso, bornIn, Spain*> et <*Pablo Picasso, bornIn, Europe*>. En d'autres termes, les faits plus généraux qu'un fait considéré comme vrai seront nécessairement, eux aussi, toujours vrais ; formellement $\forall f, f' \in F_d : f \preccurlyeq f' \wedge tf(f) \Rightarrow tf(f')$ ce qui signifie que pour $f = (d, v)$ et $f' = (d, v')$ avec $v, v' \in V_d$, on a $v \preccurlyeq v' \wedge tf(f) \Rightarrow tf(f')$. Cette définition signifie qu'un ensemble de valeurs peuvent être considérées

[5] Dans la suite, nous utilisons une *sous-ontologie* de l'ontologie associée à DBpedia – http://wiki.dbpedia.org/services-resources/ontology

comme vraies pour une description $d \in D$ particulière – on note V_d^* l'ensemble des valeurs vraies associées à la description d. Cela signifie que si une source exprime un fait, la source exprime également de façon implicite l'ensemble des faits plus généraux que le fait exprimé.

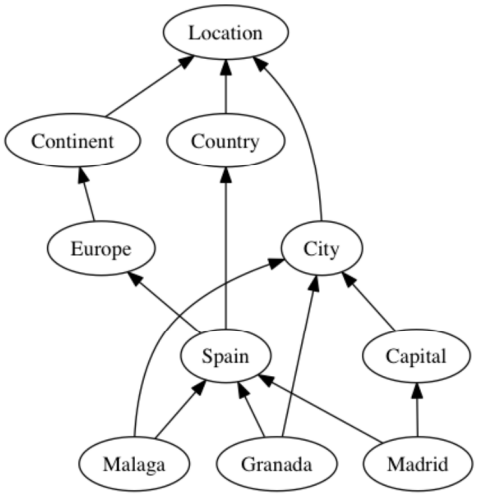


FIGURE 1 – Exemple d'un ordre partiel entre certaines valeurs, qui inclue les relations de spécialisation **`subClassOf`** et de composition **`partOf`**

Si l'on observe les contraintes qui définissent l'espace (i.e. l'ensemble) des valeurs vraies, différentes propriétés générales de $V_{d\in D}^*$ peuvent être exprimées. Ces propriétés sont fondamentales et vont être à la base de la définition de la sémantique du modèle proposé par la suite. Comme dans les approches classiques nous considérons que les faits fournis permettent à eux seuls de dériver, non plus la valeur vraie, mais l'ensemble des valeurs vraies associées à une description. Ainsi, sans connaissance supplémentaire nous considérons que l'ensemble des valeurs vraies associées à une description est inclus dans l'ensemble des valeurs induites par les valeurs V_d proposées dans les faits F_d:

$$V_d^* \subseteq \bigcup_{x \in V_d} \{y | x \preccurlyeq y\} \tag{2}$$

Nous allons cependant toujours considérer que dans l'absolu, et en accord avec la notion de prédicat fonctionnel, une valeur unique permet à elle seule de dériver l'ensemble des valeurs vraies associées à une description :

$$\forall d \in D, \exists x \in V_d^* \text{ tel que } V_d^* = \{y | x \preccurlyeq y\} \tag{3}$$

Cela implique que l'ensemble des valeurs vraies possibles V_d peut contenir des paires de valeurs qui sont incompatibles ou en conflit, i.e. des paires de valeurs qui ne peuvent pas apparaître toutes les deux dans l'ensemble de valeurs vraies associé à une description. Formellement les valeurs qui sont incompatibles sont représentées par les paires $(x, y) \in V^2$ pour lesquelles $\nexists z \in V$ tel que $\neg(x \preccurlyeq y \bigvee y \preccurlyeq x) \wedge (z \preccurlyeq x \wedge z \preccurlyeq y)$ – dans la Figure 1 *Spain* et *Capital* ne sont pas en conflit, *Malaga* et *Granada* le sont : ces valeurs ne sont pas ordonnées et il n'existe pas de valeur qui les spécialise. En s'accordant sur cela nous considérons de la connaissance non explicitée par l'ordre partiel, ici, que *Malaga* et *Granada* font référence à des localisations distinctes – sans territoire partagé. Cependant, du fait que la plupart des techniques de représentation des connaissances basées sur les logiques descriptives considèrent l'assomption d'un *monde ouvert*, deux valeurs ne peuvent être définies comme entrant en conflit que si elles sont explicitement précisées comme disjointes et si l'on sait que les deux valeurs font en effet référence à deux entités distinctes (assomption du nom unique).

Ainsi pour une description $d \in D$, en fonction des remarques précédentes et d'une valeur vraie $v \in V_d^*$, avec V_d^* inconnu, nous pouvons tout de même inférer de la connaissance sur V_d^* en excluant toutes les valeurs de V_d qui sont en conflit avec v. Néanmoins, sans connaissance supplémentaire sur V_d^*, il est impossible de s'exprimer sur l'ensemble des valeurs qui spécialisent v – dans ce contexte ces valeurs sont considérées comme étant en *conflit potentiel*. Cette relation entre valeurs n'est pas symétrique : *Granada* est en conflit potentiel avec *Spain*, alors que *Spain* est en accord avec *Granada*. Dire que quelqu'un est né à Grenade implique de dire qu'il est né en Espagne, alors que le contraire n'est naturellement pas vrai.

De façon plus générale, identifier l'ensemble des valeurs vraies pour une description donné $d \in D$ revient à identifier l'ensemble $V_d^* \subseteq V_d$ respectant les contraintes (2) et (3) qui maximisent la confiance au regard de la confiance associée aux faits de F_d qui contiennent les valeurs de V_d^*.

Adopter une telle approche nécessite la définition d'une fonction objectif permettant de calculer la fiabilité associée à une source ; cette fonction étant naturellement définie en tenant compte de l'appréciation de la confiance associée à chaque fait. Cela nécessite donc de considérer des contraintes ou de la connaissance supplémentaires par rapport aux solutions souhaitées (optimisation de deux critères dépendants). Les approches itératives sont particulièrement adaptées pour amener la résolution de ce genre de problème. Comme nous l'avons vu lors la définition de la notion de vérité (Equation 1) et lors de la définition des contraintes définies par l'ordre partiel de valeurs, des propriétés intéressantes sur l'ensemble des valeurs vraies peuvent être dérivées pour chaque description. Plus généralement, ces propriétés précisent comment l'information amenée par l'observation de faits doit être propagée dans l'objectif de distinguer les ensembles de valeurs vraies associées aux descriptions ainsi que la confiance à associer aux sources. Comme nous allons le voir dans la section suivante, de façon intéressante, la définition de l'espace des valeurs vraies que nous avons proposée répond au cadre défini par les fonctions de croyance qui sont classiquement utilisées pour traiter des données incertaines et imprécises.

3.2 Approche proposée

La modélisation de la solution proposée repose sur les fonctions de croyance introduites dans (Shafer, 1976). Ces fonctions permettent de représenter l'ignorance et l'incertitude contenues dans des informations contradictoires. Pour faciliter la lecture, nous présentons notre approche en nous appuyant sur une adaptation des notations habituelles en théorie des croyances. L'unité atomique manipulée par ces fonctions est la fonction de masse qui, dans notre cas, peut être vue comme une fonction $m_d: V \rightarrow [0,1]$ qui dépend d'une description $d \in D$ considérée. Cette fonction représente la *portion de preuve* allouée à une valeur particulière (et non pas plus spécifique). Elle peut être utilisée pour définir la croyance (*belief* en anglais) qui peut être associée à une valeur spécifique.

$$Bel_d(v) = \sum_{v' \preccurlyeq v} m_d(v') \tag{4}$$

Cette formule permet de sommer l'information apportée par l'observation d'une valeur ; elle est ainsi en totale adéquation avec la définition de l'ensemble des valeurs vraies définie plus haut. Dans notre cas, la fonction de croyance propage l'information véhiculée par un fait aux faits qui lui sont plus généraux en considérant l'ordre partiel défini par l'ontologie. La contrainte de place nous empêche de détailler certains aspects techniques de l'approche adoptée et du lien établi avec les fonctions de croyance, mais le lecteur pourra se référer à (Harispe *et al.*, 2015) pour les détails relatifs à l'utilisation de ces fonctions en considération d'un ordre partiel.

A titre illustratif, nous proposons d'adapter le modèle de découverte de vérité *Sums*, défini dans (Pasternack & Roth, 2010), en y intégrant la nouvelle formulation du problème et la prise en compte du modèle de propagation présenté. La méthode *Sums* adopte une procédure itérative dans laquelle le calcul de la fiabilité associée à une source et le calcul de la confiance associée à un fait sont alternés jusqu'à atteindre une convergence. Les formules utilisées dans la définition originale sont les suivantes :

$$t^i(\mathrm{s}) = \sum_{\mathrm{f}\in F^s} c^{i-1}(\mathrm{f}) \tag{5}$$

$$c^i(f) = \sum_{\mathrm{s}\in Sf} t^i(\mathrm{s}) \tag{6}$$

avec t^i l'estimation de la fiabilité associée à une source et c^i la confiance associée à un fait respectivement à l'itération i. Noter que l'approche itérative requiert une phase d'initialisation pour une des quantités à estimer. Dans nos expérimentations, nous avons choisi d'attribuer une même confiance à tous les faits. La fiabilité associée à une source $s \in S$ est ensuite évaluée en sommant les confiances sur les faits qui lui sont associés. De façon similaire, la confiance associée à un fait, $c^i(f)$, est évaluée en sommant les fiabilités des sources qui expriment ce fait. A chaque itération une étape de normalisation est appliquée : $t^i(\mathrm{s})$ et $c^i(f)$ sont divisés par $max_{s\in S}\left(t^i(s)\right)$ et $max_{f\in F}\left(c^i(f)\right)$ respectivement.
L'approche *Sums* peut être adaptée à notre problématique en modifiant le calcul de la confiance d'un fait. Au lieu de ne considérer que l'ensemble des sources qui expriment un fait, on va tenir compte de la transitivité de l'ordre partiel et modifier S^f par S^{f+}. On aura donc $c^i(f) = \sum_{\mathrm{s}\in Sf+} t^i(\mathrm{s})$ avec S^{f+} défini comme l'ensemble des sources qui proclament un fait donné et des sources qui proclament des faits plus spécifiques. Autrement dit, $S^{f+} = S^f \cup \{s \in S^{f'} : f' \in F \wedge f' \preccurlyeq f\}$. Notez tout de même que la façon de calculer la fiabilité d'une source ne tient pas compte de l'ordre exprimé sur les faits, pour ne pas intégrer à deux reprises la même information.
Une conséquence importante de cette modification concerne le nombre de valeurs de vérité. Ainsi l'adaptation de la méthode *Sums*, ou de toute autre méthode, nécessite la définition d'une stratégie permettant de distinguer l'ensemble des valeurs vraies après convergence. Pour cela nous utilisons un algorithme glouton (non détaillé ici par manque de place) qui répond à la stratégie suivante. L'algorithme démarre à la racine de l'ordre partiel défini sur les valeurs et dans un parcours en profondeur, cherche à maximiser la confiance des valeurs visitées à chaque itération. Ainsi à chaque itération, pour la valeur considérée, l'algorithme sélectionne parmi ses descendants directs (enfants), la valeur qui a la confiance maximale. Si cette valeur a une confiance supérieure à un seuil prédéfini qui traduit la valeur minimale de confiance admise, la procédure récursive est invoquée à nouveau. Dans le cas contraire, le programme s'arrête et la valeur v courante est considérée comme la valeur vraie la plus spécifique. A partir de cette valeur l'ensemble des valeurs vraies V_d^* est généré, tel que $V_d^* = \{y | v \preccurlyeq y\}$.

4 Evaluation de la méthode

Notre objectif étant d'adapter des méthodes existantes afin de prendre en compte la connaissance du domaine et la relation d'ordre entre les valeurs, nous avons été amenés à créer un nouveau jeu de tests car aucun de ceux proposés dans la littérature ne faisait l'hypothèse de relations possibles entre les valeurs associées à des descriptions.

4.1 Constitution du jeu de test

En effet, l'un des jeux de données les plus populaire dans ce domaine, à savoir celui des *Auteurs* présenté dans (Dong *et al.*, 2010) contient une liste d'auteurs pour un ensemble de livres. Il est clair qu'on ne peut pas avoir de relation d'ordre partiel sur ces auteurs, identifiés par leurs noms propres. La même constatation s'impose pour les jeux de données proposés par (Pasternack & Roth, 2010) qui concerne pour l'un la population (la taille de chaque ville) et pour l'autre des données biographiques (dates de naissance et de décès de personnes).

Notre objectif était de créer un jeu de test qui regroupe i) un ensemble de descriptions pour lesquelles ii) un ensemble de valeurs vraies est connu, iii) un ensemble de sources et iv) un ensemble de faits associé à chaque source.

Nous avons collecté un ensemble de faits de DBpedia (Auer *et al.*, 2007) considérés comme étant tous vrais (postulat). Nous nous sommes focalisés pour cette extraction sur le prédicat **`dbpedia-owl:birthPlace`** (version 2015-04) et nous avons choisi les faits pour lesquels il n'y avait pas de doublon. Nous avons ensuite généré des sources avec un degré de fiabilité associé. Ce degré était fixé à une valeur moyenne pour la plupart des sources et très faible pour un petit nombre d'entre-elles. Nous avons ensuite respecté les règles suivantes :

- Une source ne propose pas de fait associé à l'ensemble des descriptions vraies ;
- Une source propose un fait vrai en fonction de son degré de fiabilité et peut choisir comme valeur, un ancêtre de la valeur identifiée comme étant vraie (nous utilisons pour cela une mesure de similarité, grâce à la SML[6], afin de nous limiter dans la liste des ancêtres). Trois types de jeux de données[7] ont été définis : EXP, LOW_E et UNI qui diffèrent par la stratégie de sélection des valeurs vraies (c.f. Figure 2) ;
- Une source propose un fait faux, en fonction de son degré de non-fiabilité ($1 - degré\ de\ fiabilité$) et choisit à cet effet des valeurs qui n'ont aucun lien de généralisation/spécialisation avec la valeur vraie donnée (pour ce faire une mesure de similarité est également utilisée). Des valeurs fausses déjà proposées ont une plus grande probabilité d'être sélectionnées.

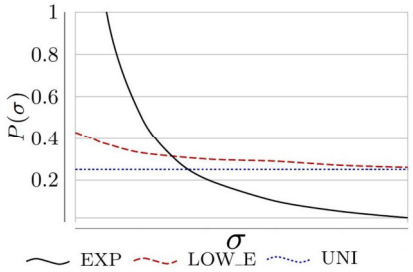


FIGURE 2 – Distributions utilisées pour la sélection des valeurs "vraies"

Basés sur ces règles, différents jeux de données ont été générés sur lesquels nous avons expérimenté notre algorithme. Pour chaque jeu de données généré, nous avons vérifié que le pourcentage de valeurs vraies et de valeurs fausses proposées par chaque source correspondait bien au degré de fiabilité qui avait été associé au préalable à chaque source.

4.2 Méthodologie d'évaluation

Pour chaque expérimentation, la valeur initiale de la confiance a été fixée arbitrairement à 0,5. Le critère d'arrêt de l'itération est le même que dans (Pasternack & Roth, 2010).

[6] Semantic Measure Library (Harispe, Ranwez, Janaqi, & Montmain, 2014)

[7] 20 jeux de données pour chacun des trois types cités.

L'algorithme a été implémenté en Python et les tests ont été réalisés sur un PC Intel Core 2 Duo processor (2.93GHz / 4.00GB). Nous ne pouvons pas baser nos évaluations sur les mesures de précision, rappel ou pertinence comme c'est souvent le cas dans la littérature. En effet, contrairement aux autres approches, nous sélectionnons un ensemble de valeurs comme pouvant être vraies. La probabilité que la valeur exacte appartienne à cet ensemble est donc supérieure. Nous avons donc préféré nous baser sur le taux d'erreurs obtenu par une méthode classique (*Sums*) et celui obtenu par son adaptation avec notre approche (*adapted_Sums*).

4.3 Résultats

Sur chaque jeu de données, *Sums* et *adapted_Sums* ont été appliqués. La Table 1 synthétise les résultats obtenus. En fonction du type de jeu de données et de la méthode de détection utilisée, le taux d'erreur moyen est présenté (moyenne calculée sur les 20 jeux de tests de chaque type). On peut y voir que la prise en compte d'une relation d'ordre partiel entre les valeurs a un impact significatif sur le taux d'erreur. En effet, celui-ci est sensiblement meilleur (plus faible) pour tous les jeux de données sur lesquels on utilise une adaptation prenant en compte l'ontologie de domaine.

TABLE 1 – *Taux d'erreur moyen pour chaque type d'évaluation.*

Dataset	Model	Error rate
UNI	*Sums*	0.269
UNI	*Adapted Sums*	**0.171**
LOW E	*Sum*	0.250
LOW E	*Adapted Sums*	**0.173**
EXP	*Sum*	0.206
EXP	*Adapted Sums*	**0.172**

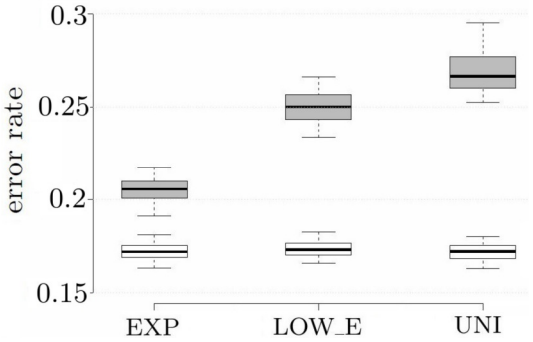


FIGURE 3 – Taux d'erreur en fonction des différents types de test : *Sums* en gris, *Adapted_Sums* en blanc

La Figure 3 montre les comparaisons pour chaque type de jeu de données entre l'application de la méthode de référence (*Sums*, boîtes grises) et la méthode adaptée tenant compte de l'ordre partiel (*adapted_Sums*, boîtes blanches). On y remarque que la nature du jeu de données (dépendant de la stratégie de sélection des valeurs vraies) n'influence pas les résultats quand la méthode tient compte de l'ontologie, contrairement à l'approche classique. On peut donc en déduire que cette approche est plus robuste. En effet, l'utilisation d'une ontologie permet de distinguer les cas où la différence entre les valeurs est seulement syntaxique et non pas sémantique.

5 Synthèse et perspectives

Cet article propose une nouvelle modélisation de la problématique de détection de vérité dans une base de faits, qui tient compte de la modélisation de la connaissance d'un domaine (ontologie). Notons que nous restons bien dans le cas de prédicats *fonctionnels*, i.e. pour lesquels il n'existe qu'une seule valeur vraie, mais où cette valeur peut être considérée à différents degrés de granularité. En effet, pour mieux répondre à des problématiques du monde réel, il est nécessaire de considérer que différentes valeurs associées à des descriptions

de certaines entités ne sont pas nécessairement concurrentes, mais peuvent traduire un certain point de vue. Ceci correspond à la plupart des contextes où une terminologie technique est utilisée. Dans ces cas, un ordre partiel peut être appliqué sur les valeurs candidates sans forcément que celles-ci entrent en conflit. Ainsi pour une entité donnée et une description qui y est rattachée, nous proposons en ensemble de valeurs vraies (valeurs non conflictuelles). Cet ensemble est construit en utilisant la propagation de *confiance*, inspirée par les approches de la théorie des croyances, appliquée à des méthodes traditionnelles (*Sums* dans cet article). Une évaluation au travers de 60 jeux de données de 3 types distincts a été menée. Les résultats montrent qu'une adaptation des méthodes traditionnelles qui intègre la prise en compte d'une structuration entre les valeurs, au travers d'une ontologie de domaine, conduit à de meilleurs résultats : le taux d'erreur est sensiblement diminué. Par ailleurs, cette approche est plus robuste, car moins sensible à la nature des jeux de données utilisés. En effet, certains jeux contenaient une proportion de valeurs vraies variable, pour refléter les cas où de nombreuses sources peuvent émettre des faits contradictoires ou pas sur certaines entités. Notre approche est basée sur une stratégie gloutonne itérative qui fournit l'ensemble des valeurs de vérité. Les jeux de données et le code source sont disponibles à *https://github.com/valentinaberetta/TDO*.

Cette étude préliminaire souligne l'apport que constitue la prise en compte de l'ordre défini entre les concepts d'une ontologie dans la détection de vérité et ouvre de nombreuses perspectives. Nous envisageons d'étudier le comportement d'autres méthodes de la littérature, lorsqu'on les adapte avec cette prise en compte afin de vérifier la flexibilité de l'approche. Ensuite, nous analyserons d'autres caractéristiques qui peuvent être intégrées à la détection de vérité. En effet, nous n'avons considéré ici que l'ordre partiel défini sur les valeurs, mais nous n'avons pas tenu compte de la sémantique associée aux concepts de l'ontologie qui pourrait être utilisée pour lisser *l'évidence* que constitue une valeur pour les autres valeurs. De même, nous n'avons pas considéré certains motifs qui peuvent être observés dans la base de données et qui peuvent renforcer ou au contraire réduire la confiance dans certaines valeurs. Ces motifs peuvent mettre en avant des cooccurrences de faits ce qui peut renforcer la confiance en certaines valeurs. Reprenons notre exemple. Si le fait qu'une personne est née en Espagne cooccure presque systématiquement avec le fait que la même personne parle espagnol, alors le fait que Pablo Picasso parle espagnol va renforcer la confiance associée au fait qu'il soit né en Espagne. Enfin, la procédure de propagation peut être modifiée. Notre approche ne considère, à l'heure actuelle, qu'une propagation ascendante inspirée par la propagation des croyances. Cette propagation peut être améliorée en y intégrant une propagation descendante, telle que la propagation des *vraisemblances* en théorie des croyances (plausibilité). L'évidence d'un fait sera alors dépendante de l'observation de faits plus génériques et de faits plus spécifiques.

Références

AUER, S., BIZER, C., KOBILAROV, G., LEHMANN, J., CYGANIAK, R., & IVES, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. In K. Aberer, K.-S. Choi, N. Noy, D., … P. Cudré-Mauroux (Eds.), *The Semantic Web, LNCS* (Vol. 4825, pp. 722–735). Springer Berlin Heidelberg.

BLANCO, L., CRESCENZI, V., MERIALDO, P., & PAPOTTI, P. (2010). Probabilistic models to reconcile complex data from inaccurate data sources. In B. Pernici (Ed.), *Advanced Information Systems Engineering: Proc. of 22nd International Conference CAiSE 2010* (pp. 83–97). Hammamet, Tunisia: Springer-Verlag.

DONG, X. L., BERTI-EQUILLE, L., HU, Y., & SRIVASTAVA, D. (2010). Global detection of complex copying relationships between sources. In E. Bertino, P. Atzeni, K. L. Tan, Y. Chen, & Y. C. Tay (Eds.), *Proc. of the VLDB Endowment* (Vol. 3, pp. 1358–1369).

DONG, X. L., BERTI-EQUILLE, L., & SRIVASTAVA, D. (2009a). Integrating conflicting data. In S. Abiteboul, T. Milo, J. Patel, & P. Rigaux (Eds.), *Proc. of the VLDB Endowment* (Vol. 2, pp. 550–561).

DONG, X. L., BERTI-EQUILLE, L., & SRIVASTAVA, D. (2009b). Truth discovery and copying detection in a dynamic world. In S. Abiteboul, T. Milo, J. Patel, & P. Rigaux (Eds.), *Proc. of the VLDB Endowment* (Vol. 2, pp. 562–573).

DONG, X. L., GABRILOVICH, E., MURPHY, K., DANG, V., HORN, W., LUGARESI, C., … ZHANG, W. (2015). Knowledge-based trust: estimating the trustworthiness of web sources. In C. Li & V. Markl (Eds.), *Proc. of the VLDB Endowment* (Vol. 8, pp. 938–949).

FOGG, B. J., & TSENG, H. (1999). The elements of computer credibility. In M. G. Williams & M. W. Altom (Eds.), *Proc. of the SIGCHI conference on Human factors in computing systems the CHI is the limit - CHI '99* (pp. 80–87). New York, New York, USA: ACM Press.

GALLAND, A., ABITEBOUL, S., MARIAN, A., & SENELLART, P. (2010). Corroborating information from disagreeing views. In *Proc. of the third ACM international conference on Web search and data mining - WSDM '10* (pp. 131–140). New York, New York, USA: ACM Press.

GIL, Y., & ARTZ, D. (2007). Towards content trust of web resources. *Web Semantics: Science, Services and Agents on the World Wide Web*, *5*(4), 227–239.

GUPTA, M., SUN, Y., & HAN, J. (2011). Trust analysis with clustering. In *Proc. of the 20th international conference companion on World Wide Web - WWW '11* (pp. 53–54). New York, USA: ACM Press.

HARISPE, S., IMOUSSATEN, A., TROUSSET, F., & MONTMAIN, J. (2015). On the consideration of a bring-to-mind model for computing the Information Content of concepts defined into ontologies. In *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1–8). IEEE.

HARISPE, S., RANWEZ, S., JANAQI, S., & MONTMAIN, J. (2014). The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics (Oxford, England)*, *30*(5), 740–742.

KELTON, K., FLEISCHMANN, K. R., & WALLACE, W. A. (2008). Trust in digital information. *Journal of the American Society for Information Science and Technology*, *59*(3), 363–374.

LI, Y., GAO, J., MENG, C., LI, Q., SU, L., ZHAO, B., … HAN, J. (2015). A Survey on Truth Discovery. *ACM SIGKDD Explorations Newsletter*, *17*(2), 1–16.

PASTERNACK, J., & ROTH, D. (2010). Knowing What to Believe (when you already know something). In *Proc. of 23rd International Conference on Computational Linguistics, COLING'10* (pp. 877–885). Stroudsburg, PA, USA: Association for Computational Linguistics.

PASTERNACK, J., & ROTH, D. (2011). Making better informed trust decisions with generalized fact-finding. In *IJCAI'11 Proc. of the Twenty-Second international joint conference on Artificial Intelligence* (pp. 2324–2329). Barcelona, Catalonia, Spain: AAAI Press.

POCHAMPALLY, R., DAS SARMA, A., DONG, X. L., MELIOU, A., & SRIVASTAVA, D. (2014). Fusing data with correlations. In *Proc. of the 2014 ACM SIGMOD international conference on Management of data - SIGMOD '14* (pp. 433–444). New York, New York, USA: ACM Press.

QI, G.-J., AGGARWAL, C. C., HAN, J., & HUANG, T. (2013). Mining collective intelligence in diverse groups. In *Proc. of the 22nd international conference on World Wide Web - WWW '13* (pp. 1041–1052). New York, USA: ACM Press.

SHAFER, G. (1976). *A Mathematical Theory of Evidence*. Princeton: Princeton University Press.

WAGUIH, D. A., & BERTI-EQUILLE, L. (2014). Truth Discovery Algorithms: An Experimental Evaluation. *arXiv:1409.6428*, 13.

WANG, X., SHENG, Q. Z., FANG, X. S., YAO, L., XU, X., & LI, X. (2015). An Integrated Bayesian Approach for Effective Multi-Truth Discovery. In *Proc. of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15* (pp. 493–502). New York, New York, USA: ACM Press.

YU, P. S. (2008). Truth Discovery with Multiple Conflicting Information Providers on the Web. *IEEE Transactions on Knowledge and Data Engineering*, *20*(6), 796–808.

ZHAO, B., RUBINSTEIN, B. I. P., GEMMELL, J., & HAN, J. (2012). A Bayesian approach to discovering truth from conflicting sources for data integration. *Proc. of the VLDB Endowment*, *5*(6)

Des primitives visuelles pour l'assistance aux échanges entre experts et ontologues

Sylvie Despres[1], Jérôme Nobécourt[1], Fanny Rigour[1]

Université Paris 13, Sorbonne Paris Cité, LIMICS, INSERM, (UMRS 1142), Sorbonne Universités, UPMC Univ Paris 06, F-93017, Bobigny, France
sylvie.despres@univ-paris13.fr, jerome.nobecourt@univ-paris13.fr, fanny.rigour@edu.univ-paris13.fr

Résumé : L'étape d'acquisition des connaissances pour construire une ontologie formelle nécessite des échanges entre les experts du domaine et le (ou les) ontologue(s) en charge de cette tâche. Au cours de cette étape, un modèle ontologique est construit de manière itérative et collaborative. La qualité des échanges entre ces acteurs est fortement dépendante du support utilisé pour représenter les connaissances. Dans ce papier, nous étudions les primitives graphiques facilitant, *via* des visualisations semi-informelles, les phases de modélisation et de formalisation afin de développer un service Web facilitant les échanges autour de ces représentations. Par conséquent, nous étudions la visualisation des connaissances contenues dans l'ontologie correspondant au côté client du service.

Mots-clés : Visualisation graphique, Modélisation, Ontologies formelles, Acquisition de connaissances

1 Introduction

La construction d'une ontologie formelle pour un domaine de spécialité est une tâche complexe et coûteuse en temps. Nous utilisons le cycle classique de construction pour ce type de ressources (Suárez-Figueroa *et al.*, 2012) : acquisition, conceptualisation, formalisation et opérationnalisation. L'étape d'acquisition des connaissances pour construire une ontologie formelle nécessite des échanges entre les experts du domaine et le (ou les) ontologue(s) en charge de cette tâche. Au cours de cette étape, un modèle ontologique est construit de manière itérative et collaborative. La qualité des échanges entre ces acteurs est fortement dépendante du support utilisé pour représenter les connaissances. En effet, les experts ne maîtrisent généralement pas les systèmes de dénotation formelle et même si certains le peuvent, ils ne conçoivent pas leur domaine exclusivement au moyen de formules logiques.

Les systèmes de visualisation constituent un support alternatif à l'acquisition des connaissances et à leur modélisation. Dans ce contexte, une représentation graphique (RG) produite par l'ontologue ou l'expert, est utilisée comme support à la modélisation. Cette dernière est semi-informelle [1], elle constitue une trace de la construction mentale du modèle qui une fois stockée sert à la capitalisation des connaissances et peut prendre la forme de dessin, de schéma, de hiérarchie, de réseau, de graphe, etc. Elle permet également l'utilisation de la langue naturelle pour commenter tout ou partie du modèle. Ces commentaires peuvent prendre la forme d'annotations textuelles, de croquis, de symboles tels que des images ou des icônes, etc. La RG permet à l'ontologue de formaliser de manière itérative les connaissances représentées tout en laissant suffisamment de liberté à l'expert pour organiser les connaissances relatives à son expertise. Le contenu de la RG évolue au cours de la collaboration entre l'expert et l'ontologue,

1. semi-informelle : les composants graphiques peuvent être utilisés sans contraintes formelles.

au fur et à mesure de l'avancement de la construction du modèle. La RG est modifiée par des opérations de transformations (ajout, suppression, modification) traduisant un changement dans le modèle et amendée par des annotations. Par conséquent, la RG permet une confrontation des idées. L'obtention d'un consensus n'est pas forcément souhaitable. Au contraire offrir plusieurs « perspectives » pour construire un modèle peut aider à la construction de ressources modulaires. Au cours de la collaboration pour la construction du modèle, différents points de vue peuvent s'exprimer ce qui peut conduire à choisir différents types de RG. Par exemple, une représentation hiérarchique sera utilisée pour visualiser l'organisation des connaissances, une représentation radiale mettra en évidence le contexte d'utilisation d'une connaissance et un chemin sera utilisé pour expliciter la chaîne de relations liant des connaissances.

Il existe des logiciels de visualisation servant à l'acquisition des connaissances permettant de construire des cartes heuristiques [2] (Eppler, 2006) et des cartes conceptuelles (Novak, 2002). Ces cartes servent essentiellement comme support aux phases de brainstorming mais ne sont pas automatiquement formalisables. L'ontologue utilise un éditeur d'ontologie, tels que Protégé, SWOOP et NeOn toolkit [3]. Plusieurs outils de visualisation de schéma RDF existent (Pietriga & Lee, 2009) et permettent à l'ontologue de visualiser les schémas RDF sous forme graphique. Dans notre cas, nous adoptons la syntaxe XML/OWL qui permet d'utiliser l'ontologie comme un arbre XML et par conséquent les outils de la XML-family pour l'explorer. La RG de triplet RDF peut être utile pour l'ontologue mais nous ne l'étudierons pas dans cet article car notre étude est centrée sur les échanges experts/ontologues.

L'ontologie utilisée dans cette étude (Despres, 2014) est modulaire et est décrite en OWL 2 DL. Le module sensoriel a été sélectionné afin de mener des expérimentations sur l'utilisabilité de ce service avec les chercheurs de l'IPBR (Institut de Recherche Paul Bocuse). Dans ce papier, nous étudions les primitives graphiques facilitant les échanges entre experts et ontologues *via* des visualisations semi-informelles lors des phases d'acquisition et de modélisation. Nous souhaitons développer un service Web Protupos pour assister ces échanges. Par conséquent, nous nous focalisons sur la visualisation des connaissances qui sont du côté client du service.

Le papier [4] est organisé en trois grandes parties : (1) Nous analysons les primitives graphiques utiles aux échanges entre ontologues et experts après avoir identifié leurs besoins et présentons un état de l'art montrant l'existence de travaux concernant la visualisation des ontologies. (2) Après avoir analysé différentes représentations graphiques susceptibles d'être utilisées pour faciliter les échanges avec les experts, nous les enrichissons en les combinant. Les résultats sont appliqués au module sensoriel de l'ontologie. (3) Nous présentons l'architecture du service Web Protupos. Enfin nous concluons avec un premier retour d'expérience.

2 Recherche de primitives graphiques

Dans ce paragraphe nous analysons les besoins des acteurs impliqués dans la construction d'ontologies, afin de sélectionner les primitives graphiques les mieux adaptées.

2. http ://www.tonybuzan.com/about/mind-mapping/
3. https ://www.w3.org/wiki/Ontology_editors
4. Les figures en HD sont disponibles sur *http ://www-limics.smbh.univ-paris13.fr/Protupos/IC2016*

2.1 Besoins des ontologues

La tâche des ontologues est d'acquérir des connaissances dans le domaine à représenter. Parmi les outils qui leur permettent de travailler, on peut trouver des fichiers annotés, des fichiers structurés ou encore des schémas réalisés par les experts. Les ontologues ont besoin de consulter, d'explorer visuellement l'ontologie pour appréhender des informations qui ne sont pas triviales (hiérarchies, rôles, chaînes de rôles), d'identifier des patrons de conception ou de détecter des erreurs de conception. Cependant, il n'est pas toujours aisé d'appréhender la structure de l'ontologie avec les outils actuellement disponibles.

2.1.1 Exemple d'outil : Protégé

Dans Protégé, une première zone d'affichage hiérarchique permet de déplier selon une profondeur variable l'arbre des concepts. Une deuxième zone affiche la définition formelle d'un concept de la hiérarchie, des « object property », des « data property » et des individus. Enfin, une troisième zone fournit des informations sur les annotations (skos, langage naturel, ...). Ces visualisations prennent la forme de boîtes où le texte est utilisé pour présenter la partie de la description OWL2 concernée. Elles sont organisées sous forme de fenêtres textuelles pouvant utiliser des caractères semi-graphiques (caractères spéciaux, tirets pour les listes) ou la mise en évidence d'une propriété (couleur, icône). Hormis la fenêtre de la hiérarchie des concepts qui peut être ou non déroulée, ces fenêtres ne sont pas paramétrables : leur contenu est automatiquement calculé en fonction des actions de l'ontologue (sélection de la souris par exemple).

Protégé offre des outils complémentaires (FIGURE 1), accessibles via des plugins (onglets), pour dessiner sous forme d'un graphe la hiérarchie à partir d'un noeud (OWLViz [5]), ou le réseau de liens entre les concepts (ontoGraf [6]). La vue est ici progressive : l'ontologue peut dynamiquement demander l'affichage d'un nouveau niveau de la hiérarchie ou une nouvelle disposition des liens du réseau. Elle ne permet cependant pas de modifier le fichier OWL, mais uniquement de visualiser les connaissances à un niveau plus ou moins macroscopique.

Ces deux représentations peuvent être exploitées sous format PDF ou excel pour dialoguer avec les experts mais ne sont pas modifiables avec ces outils. La plupart du temps ils sont échangés par mail. Les annotations de ces ressources et leurs modifications sont réalisées avec les médias disponibles (papier/crayon, tableau blanc, etc.). Les éditeurs d'ontologies ne gèrent pas les médias utilisés pour les échanges, ils doivent par conséquent être retranscrites, reformulées et formalisées.

2.1.2 Méthodes *ad hoc*

Une première approche consiste à ajouter des composants à Protégé en utilisant l'API « Protégé OWL API » [7]. Cependant, elle reste implicitement soumise à une bonne compréhension du format de l'ontologie et par conséquent au fondement logique de cette dénotation. Il en va de même pour l'utilisation des bibliothèques de manipulation des représentations OWL comme

5. http ://protegewiki.stanford.edu/wiki/OWLViz
6. http ://protegewiki.stanford.edu/wiki/OntoGraf
7. http ://protegewiki.stanford.edu/wiki/ProgrammingWithProtege

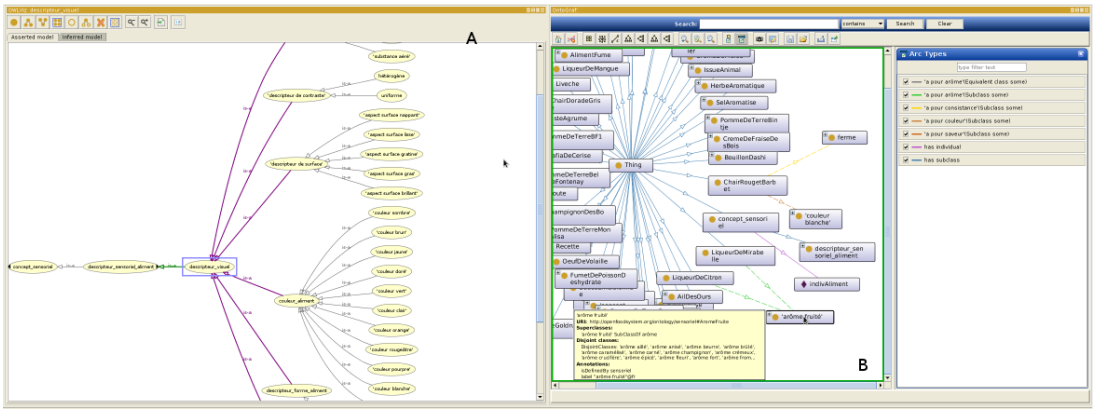


FIGURE 1 – Fenêtre A : Représentation du concept *Descripteur visuel* obtenu par OWLViz. On y voit ses ascendants et ses descendants ainsi que le type de liens le liant aux autres concepts ; Fenêtre B : Représentation du réseau autour de *Thing* obtenu par OntoGraf, affichage de certaines propriétés de concepts et annotations du concept *arôme fleuri*.

par exemple OWL API [8] et Jena [9]. Une seconde approche consiste à développer des palpeurs [10] permettant d'extraire des informations du fichier OWL en utilisant des chaînes de traitements (par exemple, XPath [11] pour extraire une partie de l'arbre OWL, SPARQL pour extraire des triplets précis de l'ontologie, etc.).

2.1.3 Vers une prise en charge des échanges entre ontologues et experts

Protégé est un outil adapté aux besoins des ontologues mais reste difficile d'accès aux utilisateurs non formés à l'aspect formel des logiques de description. Protégé est généralement perçu comme une boîte noire par les experts du domaine. Des outils comme WebProtégé (Horridge *et al.*, 2014) gèrent l'historisation destinée aux ontologues mais ne proposent pas une plateforme d'échanges entre ontologues et experts. De récents travaux (Lohmann *et al.*, 2014, 2015, 2016) visent à rendre plus accessible l'aspect formel des ontologies. Une représentation graphique du réseau des connaissances, où chaque primitive graphique décrit les opérateurs de OWL2, est présentée à l'utilisateur [12]. Il existe un état de l'art (Katifori *et al.*, 2007; Walk *et al.*, 2013) sur les besoins en visualisation des éditeurs d'ontologies. Les différents travaux présentés sont centrés sur les besoins des ontologues et n'abordent pas les échanges entre ontologues et

8. http ://owlapi.sourceforge.net/ et http ://owlcs.github.io/owlapi/

9. https ://jena.apache.org/

10. Palpeur : chaîne de traitement permettant de prendre connaissance expérimentalement de la nature d'un flux de données.

11. Comme nous pouvons le voir dans https ://www.w3.org/TR/owl2-primer/ et dans https ://www.w3.org/TR/owl2-xml-serialization/ OWL a une écriture sous forme XML. Dans le cadre de Protupos, nous préférons utiliser ce type d'écriture à une notation fonctionnelle par exemple.

12. Dans les articles sont mentionnés le terme d'expert mais sans donner plus d'information sur ce que sont ces experts. Au vue de la complexité des primitives proposées ($\cup$, $\cap$ ...) nous considérons que ce sont au minimum des experts formés aux dénotations formelles.

experts. Nous constatons qu'il y a un véritable manque dans la prise en charge des échanges entre ontologues et experts. Le service Web Protupos que nous proposons intervient dans la phase de modélisation et constitue une plateforme pour prendre en compte ces échanges.

2.2 Besoins des experts

2.2.1 Un retour d'expérience dans le cadre du projet OFS

Dans le cadre du projet OFS [13], les experts sont des chefs cuisiniers, des anthropologues, des chercheurs en science des aliments et des chercheurs en nutrition. Ils interviennent en général pour permettre d'acquérir des connaissances, valider et si nécessaire faire évoluer les modèles de connaissances relatifs aux différents modules de l'ontologie.
Les échanges se sont déroulés au cours d'ateliers centrés sur des thèmes particuliers au cours de réunions en présentiel ou par téléphone et/ou par mail. La mise en place de partage de ressources est par conséquent nécessaire. Au cours des discussions, parmi les médias utilisés, figurent : des dessins, des tableaux blancs et des captures d'écran, des fichiers au format structuré. Par exemple, les chefs peuvent avoir comme support des formats papiers représentant les captures issues de Protégé pouvant ainsi annoter les hiérarchies. Les annotations papiers/crayons permettent de garder la trace des modifications à effectuer sur le modèle ; les dessins, de préciser certaines notions ; les captures d'écran, de visualiser les relations hiérarchiques et transversales représentées dans l'ontologie. Lors de brainstorming, le tableau blanc est utilisé pour coller des post-it et ainsi créer ou modifier des modèles (regrouper par classes et liens entre les classes avec des relations).

Les échanges ont pris la forme :

— d'entretiens auprès de deux chefs (un cuisinier et un pâtissier). Ils nous ont permis la familiarisation avec les notions de base en cuisine et en pâtisserie et d'acquérir les connaissances relatives à la modélisation du domaine ;

— d'ateliers mis en place avec des chercheurs dans la domaine du sensoriel de l'IPBR. Une séance de brainstorming a permis d'identifier les premiers éléments de connaissances liés aux aspects sensoriels faisant référence au contexte du repas (commensalité), à la saveur et au mode de préparation des ingrédients, à la température (glacé, froid, tiède, chaud), à la sensation liée au plaisir (hédonisme, émotion) et à la perception (texture en bouche, aspect visuel, saveur, goût, odeur, son).

— d'ateliers qui ont eu lieu avec des anthropologues impliqués dans des familles. Les séances de brainstorming ont, là encore, permis d'identifier certains des déterminants influençant le choix d'une recette ou d'un menu. Ils sont temporels lorsqu'il s'agit des périodes de l'année où des changements d'habitude interviennent [14] ou de la saison. Le type des repas est déterminé par un moment [15] et les convives y participant [16], la composition de la famille, le coût et l'approvisionnement, le contenu du réfrigérateur/congélateur, les restes à accommoder, les matériels disponibles pour la réalisation d'un plat, le temps

13. http ://www.openfoodsystem.fr/
14. rentrées scolaires - vacances
15. semaine/WE/repas de fête
16. individu, famille/avec invité

disponible et la durée de la recette. Cela a donné lieu à la construction de plusieurs cartes conceptuelles.

L'objectif de ces échanges consistait à aboutir à une modélisation soit en validant les choix de l'ontologue une fois ces derniers explicités, soit en proposant une nouvelle représentation des connaissances. Nous avons conclu de ces expériences que les experts ont besoin de visualiser l'ontologie et d'appréhender sa structure pour pouvoir donner leur avis (accord sur la structure, ajout de sous hiérarchie, déplacement d'une partie de la hiérarchie). Actuellement, dans le projet OFS, les experts expriment leurs opinions en amendant manuellement le modèle papier et les fichiers structurés.

Les échanges entre experts et ontologues se font plus facilement *via* des représentations graphiques puisqu'elles permettent une meilleure appropriation de la conceptualisation (Quillian, 1968; Kayser, 1997). Un certain nombre d'actions sur ces représentations graphiques, telles que : ajouter, déplacer, modifier, supprimer un concept ou une arborescence et effacer (revenir en arrière) doivent également être réalisables. L'expert et l'ontologue doivent pouvoir annoter et commenter afin d'exprimer au mieux leurs idées.

2.2.2 Synthèse des besoins

Les fonctionnalités de visualisation de l'ontologie ou de ses parties constituent un moyen d'assister les interactions entre ontologues et experts. L'expert doit pouvoir visualiser les propositions des ontologues et échanger avec eux de manière synchrone ou asynchrone. Il ressort de notre analyse un certain nombre de besoins :

— appréhender la structure globale de l'ontologie (B1) ;
— contextualiser (B2) ;
— détecter des erreurs de conception (B3) ;
— naviguer dans l'ontologie (B4) ;
— visualiser des relations hiérarchiques (B5) ;
— visualiser des rôles ou chaîne de rôle (B6) ;
— visualiser l'environnement d'un concept (B7).

Notre objectif est de déterminer les primitives graphiques les mieux adaptées pour chaque acteur afin d'obtenir des échanges productifs.

3 Vers la spécification du service Web Protupos

Le service Protupos doit offrir différentes manières d'annoter les visualisations. Ainsi il doit être possible d'utiliser le langage naturel pour commenter, annoter en utilisant une palette de couleur conforme aux habitudes de l'expert et dessiner sur la RG. En outre, tous les acteurs doivent pouvoir accéder à l'ensemble des annotations. Une fois ces fonctionnalités déterminées, nous avons étudié les travaux existants. Il existe des études au niveau de la représentation des données (Héon *et al.*, 2009, 2010) et de la visualisation de l'information (Alsallakh *et al.*, 2014). Ces travaux nous ont permis d'appréhender ce qu'il est possible de faire quand on se situe au niveau des données. Ils doivent être adaptés au niveau connaissance « Knowledge Level » (Newell, 1982) car c'est le cœur de notre approche.

Nous avons retenu deux types de primitives graphiques. Les unes concernent l'organisation de la visualisation (FIGURE 2) et les autres concernant les actions potentiellement réalisables

sur ces représentations (FIGURE 3).

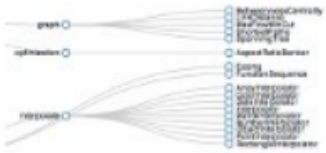
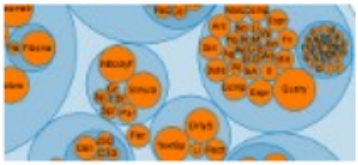
Primitives graphiques	Utilisation	Répond aux besoins :
 (1) : Cluster	Principalement utilisé pour visualiser des dendrogrammes. Les regroupements se font généralement sur des liens de type hiérarchique et méronymique comme : « kind of » et « part of »	B5 : Pour tous B3 : Ontologue
 (2) : Pack	Permet une visualisation des feuilles de l'ontologie. Son avantage est qu'elle permet une vision en coupe transversale montrant les feuilles et leurs ascendants visibles	B1 : Pour tous B3 : Ontologue B7 : Pour tous
 (3) : Partition	Donne une vision radiale du voisinage du concept sélectionné. Elle permet de voir par décomposition de couches et lorsqu'elle est interactive permet de parcourir l'ontologie	B1 : Pour tous B7 : Pour tous
 (4) : Force	Particulièrement intéressante pour se focaliser sur un concept ou sur un type de propriété et voir le réseau. Elle permet aussi de mettre en évidence les chaînes de propriétés. En effet grâce au réseau, il est possible de souligner une chaîne de propriétés en la mettant en surbrillance	B6 : Ontologue B7 : Pour tous

FIGURE 2 – Les types d'organisation retenues (cluster, pack, partition, force) et besoins associés. [Pour tous = Expert et Ontologue]

La combinaison de certaines de ces organisations et actions, permet d'obtenir des graphiques plus riches et plus informants. Actuellement, nous avons sélectionné deux graphiques combinant ces primitives :

— *(1)+(a)+(b)* : nous obtenons une vue permettant une meilleure visualisation des clusters ;
— *(1)+(3)+(a)+(b)+(c)* : nous obtenons une vue hiérarchique présentant les regroupements par secteur et qui s'auto-adapte au clic souris.

La FIGURE 4 présente ces deux modes de visualisations appliquées au module sensoriel de l'ontologie. Le graphique (1)+(a)+(b) permet de mieux percevoir la densité de certains clusters et l'importance d'un concept par rapport aux autres. Nous souhaitons l'enrichir en donnant aux acteurs la possibilité de zoomer sur un concept particulier comme par exemple, le descripteur d'aspect et le descripteur de couleur. Actuellement, cette représentation ne permet pas de parcourir interactivement l'ontologie. Tandis que le graphique(1)+(3)+(a)+(b)+(c) offre cette possibilité. Pour le module sensoriel il est possible de se déplacer à partir du concept descripteur

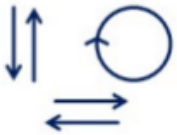
Primitives graphiques	Utilisation	Répond aux besoins :
 (a) : Hiérarchie	Consiste à afficher la visualisation niveaux par niveaux	B5 : Pour tous
 (b) : Orientée	En plus des 4 directions cardinales, la visualisation radiale permet une distribution sur 360 degrés	B4 : Pour tous (selon un axe)
 (c) : Dynamique	Ce type d'organisation permet d'afficher interactivement les éléments du modèle de manière progressive	B4 : Pour tous (utilisateur proactif)
 (d) : Personnalisable	Utilisation de formes, de couleur et de différents niveaux de zoom en fonction des besoins	B2 : Pour tous

FIGURE 3 – Les types d'actions retenues (hiérarchie, orientée, dynamique, personnalisable) et besoins associés. [Pour tous = Expert et Ontologue]

de couleur et de remonter au père qui est un descripteur d'aspect (FIGURE 5).

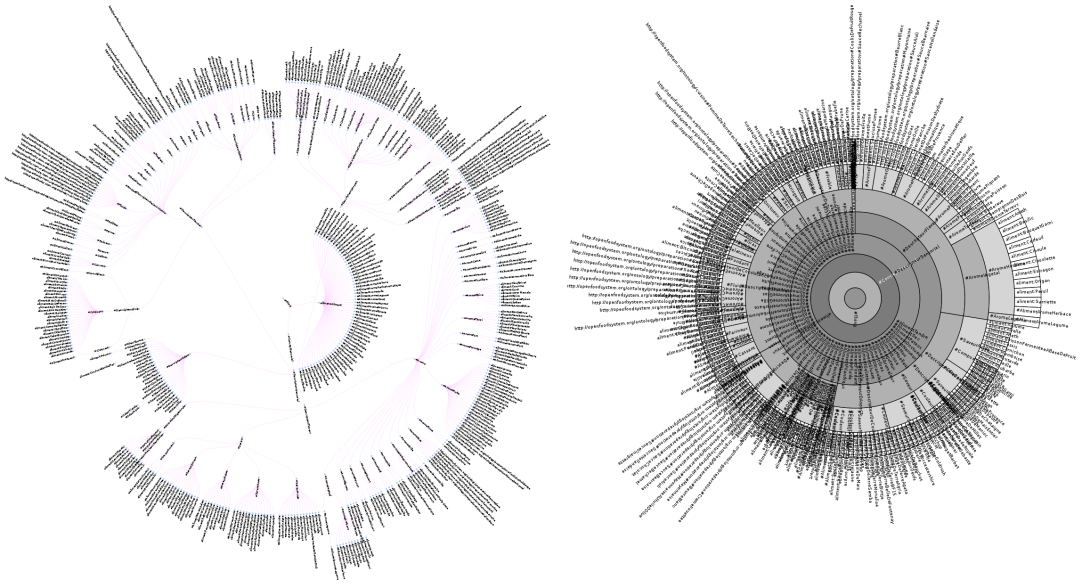


FIGURE 4 – Représentation du module sensoriel suivant deux types de vues. À gauche : (1)+(a)+(b). À droite : (1)+(3)+(a)+(b)+(c).

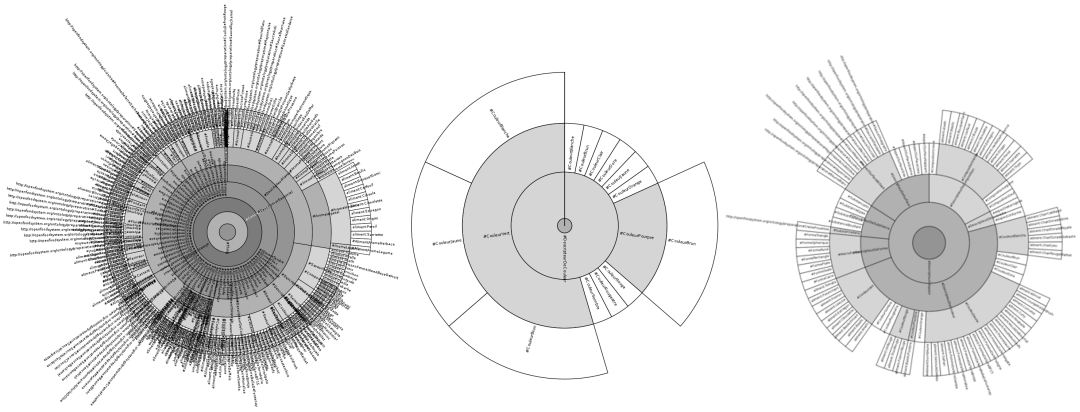


FIGURE 5 – Représentation du module sensoriel suivant le second mode. La figure est très dense et difficile à lire mais elle permet d'accéder d'un niveau 0 à n'importe quel niveau n plus spécifique, et de ce niveau il est possible de remonter.

Dans Protupos, nous souhaitons pouvoir naviguer entre les différentes RG. Il nous semble important de pouvoir changer le point de vue (hiérarchie, relation avec d'autres concepts...) et la nature de la RG au moment où nous en avons besoin. Par exemple, dans une visualisation en Partition (3) nous souhaitons explorer une Hiérarchie (a) avec une Force (4) des propriétés ce qui peut mettre en évidence les chaînes de propriétés.

4 Protupos

Protupos est un service Web (Despres & Nobécourt, 2016) qui respecte les « bonnes pratiques » du W3C [17]. Son architecture est fondée sur une organisation mixte : PHP et OWL/XML côté serveur, JS, XML et SVG côté client. L'application est servie en XHTML.

Le côté serveur sert à traiter les ressources ontologiques (fichiers owl), gérer l'historisation, calculer la visualisation demandée et les sessions permettant d'identifier les acteurs et configurer leur espace de travail. Du côté client, une attention particulière a été mise sur l'accessibilité (media queries CSS3, WAI-ARIA), la visualisation sous la forme d'objet (notion de groupe de SVG), et l'adaptation de ces visualisations aux préférences des différents acteurs.

Protupos suit une conception modulaire. Il est pensé pour pouvoir ajouter de nouveaux comportements/fonctionnalités suivant les besoins des utilisateurs. Par exemple :

— gérer la A-Box d'une logique de description ;
— traiter la composante lexicale : utilisation d'un vocabulaire contrôlé en SKOS, gestion des labels et des annotations y référant.

Le service ne remplace pas des outils comme Protégé, OWLAPI ou des commandes utilisant le shell et la XML-family. Il doit au contraire être vu comme complémentaire pour l'ontologue car il permet d'appréhender des notions complexes telles que les propriétés calculées. Ces dernières peuvent être traduites en utilisant un formalisme dédié (langages à bases de règles). Il peut également arriver qu'il ne soit pas possible de les opérationnaliser (décidabilité). Pour les experts, il permet de rendre accessible et compréhensible sous forme d'une visualisation le travail de l'ontologue. Experts et ontologues peuvent alors interagir sur cette visualisation.

Protupos permet un accès à des visualisations lorsque l'expert est disponible (service en ligne) et *via* le type de médias qu'il souhaite (tablette, ordinateur, smartphone). Ainsi des échanges *via* le Web pourront avoir lieu de manière asynchrone ce qui permettra d'augmenter la fréquence des collaborations et par conséquent d'améliorer les échanges relatifs à la construction de l'ontologie. Un tel service Web constitue le moyen de favoriser et d'améliorer la collaboration entre experts et ontologues.

Actuellement, nous débutons une phase de tests portant sur l'utilisabilté d'un tel service auprès des experts du domaine sensoriel [18]. Un premier entretien a été réalisé. Les deux visualisations présentées ci-dessus ont été testées par un expert. C'est la deuxième visualisation [(1)+(3)+(a)+(b)+(c)] qui a suscité le plus d'intérêt. Le fait que celle-ci soit interactive et que l'expert puisse parcourir le module sur tous les niveaux a été très apprécié et semble faciliter l'appréhension de la structure du module. Lors des échanges nous avons compris qu'il serait intéressant pour l'expert de pouvoir naviguer au travers des différents modules [19]. Certaines propriétés ergonomiques ont été évoquées comme la taille et la position du nom du concept qui ne sont pas encore adaptées ou encore le déclenchement de l'action sur le graphe (cliquer sur le label ou sur la section du camembert concerné). A l'issue de cette expérimentation, il apparaît

17. https ://www.w3.org/TR/mwabp/

18. Le module sensoriel décrit l'ensemble des sensations perçues à partir de la bouche : goûts (saveurs) et odeurs (arômes) mêlés, sans distinction.

19. Voilà un exemple d'échange avec l'expert : « ... par exemple, lors d'une création de recette, si le cuisiner veut ajouter un aliment croquant, il peut *via* le module sensoriel, trouver le concept "croquant" et voir les concepts faisant partie du module Aliment qui sont décrit comme "croquant". Ainsi il peut voir toutes les possibilités d'ajouter cette texture à sa recette... »

que les tests réalisés avec l'outil conduise à de nouvelles représentations répondant à ces nouveaux besoins qui devront être intégrés à Protupos. Le fait visualiser et de manipuler le module laisse libre cours à la créativité de l'expert. De nouvelles évaluations sont prévues courant mai avec de nouveaux experts de l'IPBR et des ontologues participant à d'autres projets.

5 Conclusion et perspectives

Dans cette étude, nous analysons les besoins des ontologues et des experts pour la construction d'ontologies et nous sélectionnons les types de primitives graphiques adaptées à ces besoins. Puis nous montrons que la combinaison de certaines de ces primitives graphiques conduit à des RG plus riches. Dans ce papier, deux d'entre elles ont été retenues pour tester la faisabilité [(1)+(a)+(b) et (1)+(3)+(a)+(b)+(c)]. Ensuite, nous présentons les fonctionalités d'assistance aux échanges entre les experts et les ontologues de Protupos. Enfin, nous présentons les premiers retours d'une évaluation avec un expert du domaine du sensoriel.

Dans les différentes phases de tests lors de la réalisation de ce service Web, nous avons déjà pu constater plusieurs points qui nous semble prometteurs :

— visualiser des connaissances difficilement accessibles par d'autres outils (le nombre de niveau, la complexité d'un niveau, sa granularité) ;
— faciliter la navigation dans les connaissances notamment avec des RG dynamiques ;
— visualiser des structurations de connaissances et détecter des différences dans la conceptualisation et la granularité de leurs descriptions.

Du point de vue de l'ontologue, nous constatons une véritable aide pour :

— tester progressivement la structure de l'ontologie ;
— historiser ses évolutions ;
— favoriser sa conception.

Une nouvelle évaluation sera effectuée pour tester Protupos en ligne afin d'auditer de manière plus large les services rendus.

Références

ALSALLAKH B., MICALLEF L., AIGNER W., HAUSER H., MIKSCH S. & RODGERS P. (2014). Visualizing sets and set-typed data : State-of-the-art and future challenges. In *Proceedings, 16th Eurographics Conference on Visualization*.

DESPRES S. (2014). Construction d'une ontologie modulaire pour l'univers de la cuisine numérique. In C. FARON-ZUCKER, Ed., *IC - 25èmes Journées francophones d'Ingénierie des Connaissances*, p. 27–38, Clermont-Ferrand, France. Session 1 : Construction, peuplement et exploitation d'ontologies.

DESPRES S. & NOBÉCOURT J. (2016). Protupos : vers un service web de représentations graphiques. In *Visualisation d'informations, interaction et fouille de données (Ateliers de la conférence EGC'2016)*, p. 7–8.

EPPLER M. J. (2006). A comparison between concept maps, mind maps, conceptual diagrams, and visual metaphors as complementary tools for knowledge construction and sharing. *Information Visualization*, **5**(3), 202–210.

HÉON M., BASQUE J. & PAQUETTE G. (2010). Validation de la sémantique d'un modèle semi-formel de connaissances avec ontoCASE.

HÉON M., PAQUETTE G. & BASQUE J. (2009). Méthodologie assistée de conception d'une ontologie à partir d'une conceptualisation consensuelle semi-formelle.

HORRIDGE M., TUDORACHE T., NUYLAS C., VENDETTI J., NOY N. F. & MUSEN M. A. (2014). Webprotégé : a collaborative web-based platform for editing biomedical ontologies. *Bio Informatics, Applications Note, Databases and ontologies*, **30**(16), 2384–2385.

KATIFORI A., HALATSIS C., LEPOURAS G., VASSILAKIS C. & GIANNOPOULOU E. (2007). Ontology visualization methods—a survey. *ACM Computing Surveys (CSUR)*, **39**(4).

KAYSER D. (1997). *La représentation des connaissances*. Hermès.

LOHMANN S., LINK V., MARBACH E. & NEGRU S. (2015). WebVOWL : Web-based visualization of ontologies. In *Proceedings of EKAW 2014 Satellite Events*, volume 8982 of *LNAI*, p. 154–158 : Springer.

LOHMANN S., NEGRU S., HAAG F. & ERTL T. (2014). VOWL 2 : User-oriented visualization of ontologies. In K. JANOWICZ, S. SCHLOBACH, P. LAMBRIX & E. HYVÖNEN, Eds., *EKAW*, volume 8876 of *Lecture Notes in Computer Science*, p. 266–281 : Springer.

LOHMANN S., NEGRU S., HAAG F. & ERTL T. (2016). Visualizing ontologies with VOWL. *Semantic Web*, **7**(4). To appear ; already accepted papers for the EKAW 2014 special issue.

NEWELL A. (1982). The knowledge level. *ai*, **18**, 87–127.

NOVAK J. (2002). *The Theory Underlying Concept Maps and How To Construct Them*. Rapport interne, IHMC. rev 01-2008.

PIETRIGA E. & LEE R. (2009). Langages et outils pour la visualisation et la manipulation de données du web sémantique. *Technique et Science Informatiques (TSI)*, **28**(2), 173–197.

QUILLIAN M. R. (1968). Semantic memory. In M. MINSKY, Ed., *Semantic Information Processing*, p. 227–270. Cambridge, MA : MIT Press.

SUÁREZ-FIGUEROA M. C., GÓMEZ-PÉREZ A. & FERNÁNDEZ-LÓPEZ M. (2012). The neon methodology for ontology engineering. In *Ontology Engineering in a Networked World.*, p. 9–34.

WALK S., PÖSCHKO J., STROHMAIER M., ANDREWS K., TUDORACHE T., NOY N. F., NYULAS C. & MUSEN M. A. (2013). PragmatiX : An interactive tool for visualizing the creation process behind collaboratively engineered ontologies. *Int. J. Semantic Web Inf. Syst*, **9**(1), 45–78.

Représentation systémique multi-échelle des processus biologiques de la bactérie

Vincent Henry[1], Arnaud Ferré[1], Christine Froidevaux[1], Anne Goelzer[2], Vincent Fromion[2], Sarah Cohen-Boulakia[1], Sandra Dérozier[2], Marc Dinh[2], Ghislain Fiévet[1], Stephan Fischer[2], Jean-François Gibrat[2], Valentin Loux[2], Sabine Peres[1,2]

[1] BioInfo, LRI, CNRS UMR 8623, Université Paris Sud, Université Paris-Saclay, France
{prénom.nom}@lri.fr
[2] MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France
{prénom.nom}@jouy.inra.fr

Résumé : La production à haut-débit de données biologiques de nature hétérogène nécessite une exploitation et une intégration particulières de celles-ci. Malgré le développement de nombreuses bio-ontologies, l'organisation de ces données dans un cadre structuré et adaptatif reste perfectible. Nous émettons l'hypothèse qu'une approche systémique multi-échelle de la représentation des processus cellulaires permettrait de progresser dans cette problématique. Pour valider cette démarche, nous avons conçu une modélisation ontologique des processus bactériens nécessaires à l'expression génique. Les relations entre ces processus et leurs molécules participantes ou leurs sous-processus ainsi que leurs modèles ont été formellement décrites. Cette description s'accompagne d'axiomes et de relations supplémentaires sur lesquels un raisonnement automatique est effectué. La représentation des processus réalisée permet leur mise en relation avec leurs modèles et paramètres par inférence. Parallèlement, le raisonnement apporte de nouvelles informations contextuelles de séquentialité, agrégation ou compétition. Notre contribution s'appuie sur les bio-ontologies existantes pour une meilleure interopérabilité.

Mots-clés *: Conception d'ontologie, modélisation multi-échelle, biologie systémique, raisonnement, processus cellulaire*.

1 Introduction

Les progrès constants dans les technologies à haut-débit, dites « omiques » (génomique, transcriptomique, métabolomique,…) ont permis de franchir une étape critique dans la production de données biologiques (Metzker, 2010). Cette révolution a des implications sur la masse et l'hétérogénéité des types de ces données à gérer à différentes échelles (moléculaire, cellulaire, phénotypique, *etc* ; Nekrutenko & Taylor, 2012). Ainsi, l'interprétation de la biologie à travers les sciences omiques nécessite la gestion, l'intégration et le partage de données de nature diverse. Dans ce contexte, il s'avère nécessaire de relier ces données aux mécanismes biologiques sous-jacents dans un cadre structuré et adaptatif pour conserver et analyser au mieux l'information.

Cette problématique a été identifiée dans le cadre du projet interdisciplinaire de l'Institut de Modélisation des Systèmes Vivants (IMSV) qui regroupe des biologistes, des modélisateurs et des informaticiens autour d'une question commune relative à l'intégration systémique des organismes. Afin d'y répondre, nous avons avancé l'hypothèse qu'une approche systémique multi-échelle d'un système d'organisation des connaissances (SOC) en biologie permettrait de progresser vers un modèle utile à l'intégration et l'exploitation de données hétérogènes. En effet, en biologie systémique, les données issues d'expériences omiques permettent de caractériser les paramètres de modèles représentant les processus biologiques. Ainsi ces

données, bien qu'hétérogènes, sont organisées intrinsèquement en fonction des processus biologiques et intégrées à travers les modèles mathématiques capables de les simuler et leurs paramètres associés.

L'approche systémique multi-échelle consiste à représenter chaque phénomène à l'échelle la plus pertinente. Cette représentation est rendue possible grâce à l'intégration des connaissances à différentes échelles, dans des briques (ou systèmes) élémentaires qu'il s'agit d'assembler et de manipuler de façon modulaire. L'emboîtement de ces briques offre une vue globale du système et la possibilité de zoomer sur des zones particulières permettant de représenter le comportement du système à différentes granularités. Depuis 2001, il est acquis que les fonctions cellulaires majeures peuvent être représentées par des systèmes (Kitano, 2001). S'appuyant sur ce principe, le comportement général des bactéries a pu être modélisé et simulé à différents niveaux (Goelzer & Fromion, 2011 ; Karr *et al.*, 2012 ; Goelzer *et al.*, 2015). De plus, les processus biologiques se prêtent naturellement à une représentation systémique multi-échelle, qui à son tour peut être organisée sous forme ontologique de manière méthodique (figure 1). Avec *Cell Molecular ONtology* (C'MON), nous nous proposons de tester la faisabilité et la cohérence d'une représentation ontologique des processus en lien avec leurs paramètres à partir de leur définition fonctionnelle biologique.

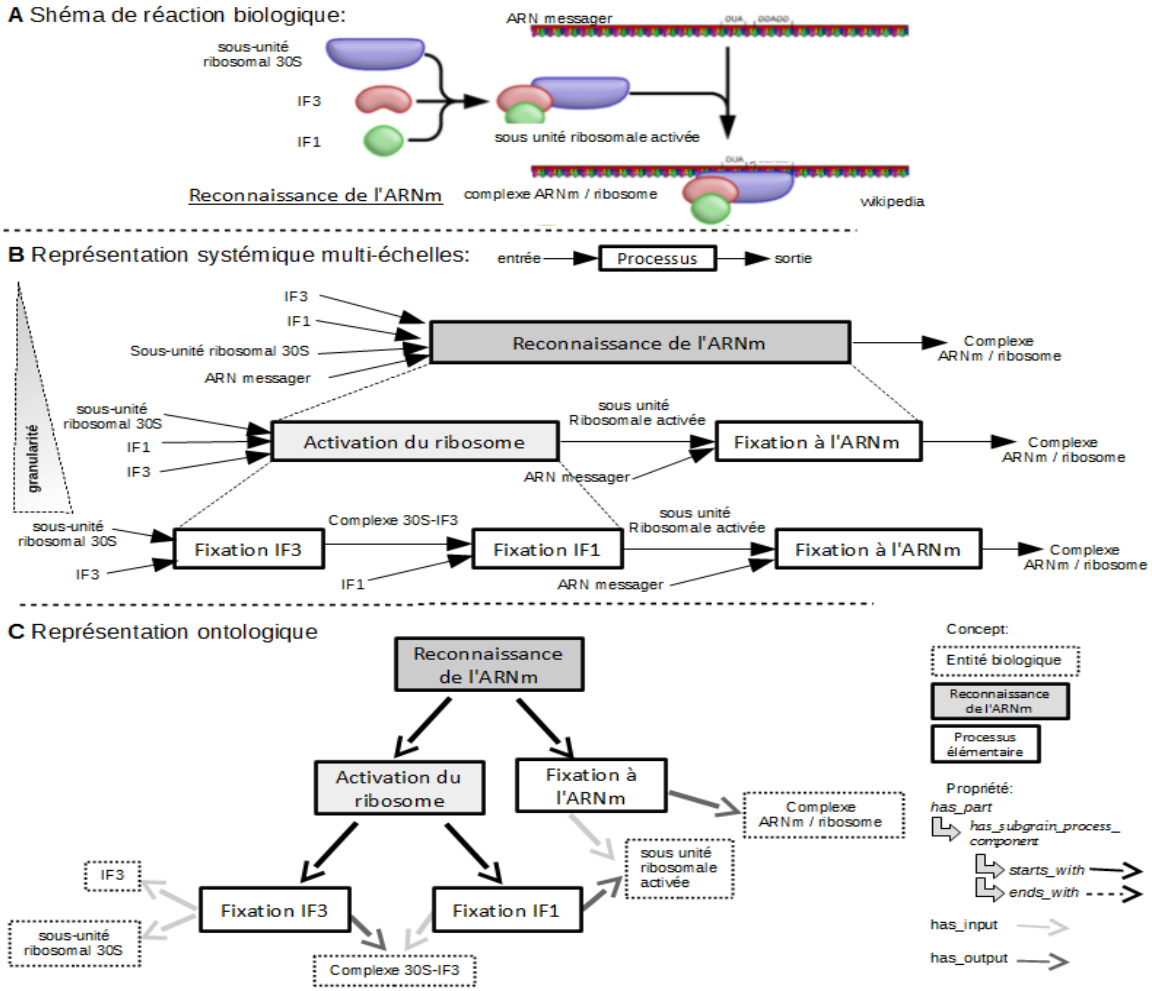


FIGURE 1 – Exemple de représentation biologique schématique, systémique multi-échelle et ontologique correspondant à un même processus cellulaire : la reconnaissance de l'ARN par le ribosome.

Des SOC existent déjà dans le domaine de la Biologie. Des bio-ontologies ont été conçues pour : l'annotation fonctionnelle des gènes et de leurs produits (*Gene Ontology* GO , GO

Consortium, 2000) ; le recensement des molécules d'intérêt biologique et de leur modification chimique en fonction de leur structure (*Chemical Entities of Biological Interest* ChEBI, De Matos *et al.*, 2010) ; la classification des séquences biologiques en fonction de leur implication dans des processus biologiques (*Sequence Ontology* SO, Eilbeck *et al.*, 2005) ; ou la hiérarchisation des principaux modèles de réactions biologiques et des paramètres associés (*Systems Biology Ontology* ; SBO ; Courtot *et al.*, 2011).

De même, des bases de connaissances comme les bases *Kyoto Encyclopedia of Genes and Genome* (KEGG) ou *Metabolic Pathway Database* (MetaCyc, Altman *et al.*, 2013) décrivent finement les réseaux métaboliques : des voies principales aux réactions enzymatiques élémentaires. Ces bases apportent aussi des informations globales et structurées sur les principaux processus cellulaires (transcription, réplication, traduction, …) et les complexes protéiques impliqués.

L'ensemble de ces SOC propose une vision assez large et relativement complète de la biologie moléculaire. Cependant, ces travaux sont développés indépendamment les uns des autres avec des objectifs différents. Néanmoins, le développement de GO-plus a démontré la possibilité d'intégrer au moins deux de ces bio-ontologies standards, GO et ChEBI (Hill *et al.*, 2013). Par ailleurs, leur étude approfondie révèle un déséquilibre d'intérêt en faveur des eucaryotes. Enfin, l'hétérogénéité des processus cellulaires et la nouveauté des connaissances fines dans ces domaines n'ont pas permis une représentation aussi précise que celle des réactions enzymatiques métaboliques.

Si aucun de ces SOC ne répond à lui seul à notre besoin, ils représentent une source de connaissance conséquente et sont assidûment utilisés en biologie. Afin de profiter au maximum de cette base de travail et de son implantation, nous attachons un soin particulier à les intégrer. A titre de preuve de concept de la possibilité de cette intégration avec une approche systémique multi-échelle, nous présentons dans cette article un modèle ontologique des processus cellulaires bactériens nécessaires à l'expression génique, C'MON.

2 Modèle ontologique

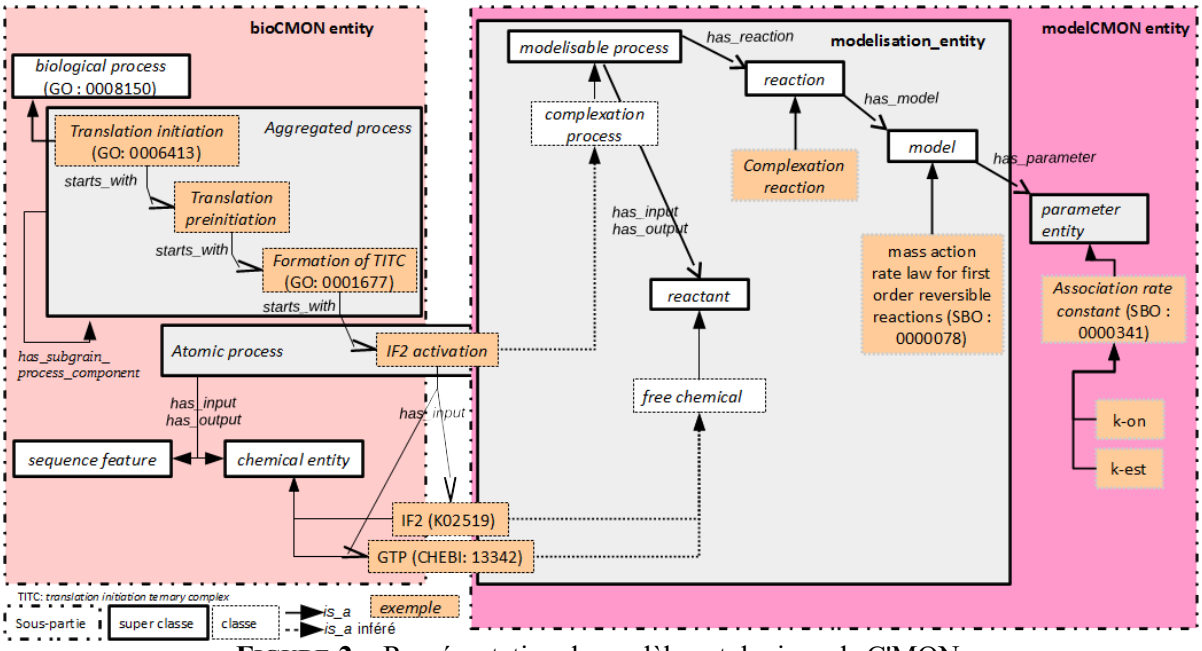


FIGURE 2 – Représentation du modèle ontologique de C'MON.

C'MON est éditée sur Protégé 5.0.0. C'MON contient 2 parties indépendantes (figure 2) : '*bioCMON entity*' (bioCMON) et '*modelCMON entity*' (modelCMON) reliées par une racine commune. bioCMON organise les processus cellulaires atomiques finement définis en fonction

de leurs entités biologiques ou les processus agrégés, hiérarchisés entre eux en sous-propriétés de *has_part.* modelCMON structure les types de paramètres reflétant les données biologiques en fonction de leur modèle. Ceux-ci sont reliés à des processus modélisables définis en fonction de leur type de réactif (reflétant les entités biologiques de bioCMON). Ainsi les processus cellulaires concepts de bioCMON sont automatiquement réorganisés dans modelCMON comme *is_a* des processus modélisables. Ils se retrouvent ainsi reliés aux modèles mathématiques qui les représentent et aux types de paramètres associés.

2.1 bioCMON

bioCMON a été conçue à partir de connaissances biologiques issues de schémas de réaction compilés par les experts biologistes de l'IMSV (figure 1A). Les informations de processus et d'entités biologiques sont extraites en suivant l'approche systémique multi-échelle. Cette étape permet d'obtenir une représentation standardisée, granulaire et centrée sur les processus encadrés par les entités biologiques en entrée et sortie (figure 1B). Cette représentation est ensuite modélisée sous forme ontologique (figure 1C). Un processus de faible granularité, agrégé (en grisé) est formellement défini par ses processus de granularité supérieure par une sous-propriété non transitive de *has_part : has_subgrain_process_component*. Celle-ci est spécifiée par des sous-propriétés *starts_with, ends_with* ou *has_intermediate*. Ces dernières reflètent les changements de granularité et apportent une information plus précise sur l'ordre relatif de réalisation des processus de granularité supérieure. Les processus de granularité maximum, élémentaires ou atomiques (en blanc) sont définis par des propriétés d'entrée(s)/sortie(s) (*input_of, output_of*) des entités biologiques (molécules, complexes macromoléculaires ou séquences) nécessaires à la réalisation du processus ou résultant de son traitement (figures 2 et 3A*).

Dans un souci d'interopérabilité, après l'extraction des connaissances, les concepts ont été prioritairement importés de bio-ontologies pré-existantes permettant la conservation des labels, *Internationalized Resource Identifier* (IRI) et identifiants (Id). Ainsi, les concepts de processus cellulaires bactériens participant à l'expression génique décrits dans GO de leur plus haute hiérarchie à la plus fine possible reliés par des *part_of* ont été importés. Les *part_of* ont été spécialisés en sous-propriétés *starts_with*, *ends_with* ou *has_intermediate*. De même, les concepts de molécules et séquences participant aux processus cellulaires et décrites dans ChEBI ou SO ont été importés. Les concepts correspondant aux complexes macromoléculaires d'intérêt décrits dans KEGG ont été créés et annotés avec l'Id correspondant grâce à la propriété standard *hasDbXref.* Les concepts d'entités biologiques ont été reliés aux processus cellulaires par des relations *input_of* ou *output_of.* Ensuite, les concepts des processus manquants par rapport au modèle systémique et leurs participants ont été ajoutés suite à un effort de curation. Si nécessaire, les processus de plus bas niveau ont été segmentés par l'ajout de nouveaux concepts de processus jusqu'à ce qu'ils impliquent au maximum trois entités biologiques (processus élémentaires).

Au final, bioCMON contient trois super-classes disjointes : '*biological process'* (contenant 40 % de concepts communs avec GO, principalement aux niveaux agrégés) qui hiérarchise les concepts de processus, '*chemical entity*' (contenant 20 % de concepts communs avec ChEBI et 15 % avec KEGG) qui hiérarchise les concepts de molécule et où les classes de mêmes parents sont disjointes, et '*sequence feature'* (contenant 85 % de concepts communs avec SO) qui hiérarchise les concepts de séquences biologiques (figure 2).

2.2 modelCMON

Les concepts modelCMON sont formalisés en lien avec bioCMON. modelCMON comprend 4 super-classes dans '*modelisation entity*' (figure 2):

i) '*reaction*' : *reaction* classifie dix réactions modélisables et est défini en relation avec la super-classe '*model*' par la relation *has_model*. Ces réactions ont été choisies car elles présentent toutes des spécificités en nombre ou en nature (libres, liés ou à motifs) de leurs réactifs.

ii) '*reactant*' : *reactant* est une reclassification des classes de bas niveau de *chemical entity* et *sequence feature* sous forme de réactifs. Ces classes sont automatiquement sélectionnées telles que :

```
reactant ≡ (input_of some 'biological process' or output_of some 'biological process')
```

En fonction de leur nature (séquence ou molécule) et de leur rôle biologique, les *reactant* sont automatiquement répartis entre réactifs libres, réactif à motifs et réactifs liés.

iii) '*modelisable process*' : les processus modélisables sont formellement décrits en fonction du nombre et de la nature de ces réactifs (figure 3A** : *complexation model*). Cette superclasse est donc une nouvelle hiérarchisation des classes de *biological process* qui peuvent être reliées à des réactions modélisables par inférence. Il existe 10 processus modélisables en lien avec les 10 réactions par la relation *has_reaction* (exemple figure 3A*** : *Complexation*).

iv) '*model*' : les modèles hiérarchisés dans cette super-classe ont été sélectionnés et intégrés à partir de SBO. On y retrouve les modèles capables de modéliser les 10 réactions. model est défini en relation avec la super-classe '*parameter entity*' par la relation *has_parameter*. *parameter entity* contient une liste de paramètres intégrés à partir de SBO.

Suite à cette formalisation, modelCMON est construit par inférence à partir de bioCMON. Le raisonneur (HermiT 1.3.8 ; Glimm *et al.*, 2014) s'appuie sur la sélection de réactifs d'intérêt puis des processus biologiques simulables. Sur ce principe, les axiomes sont utilisés pour faire le lien entre un contexte biologique et un contexte de modélisation pour une même connaissance (figure 2).

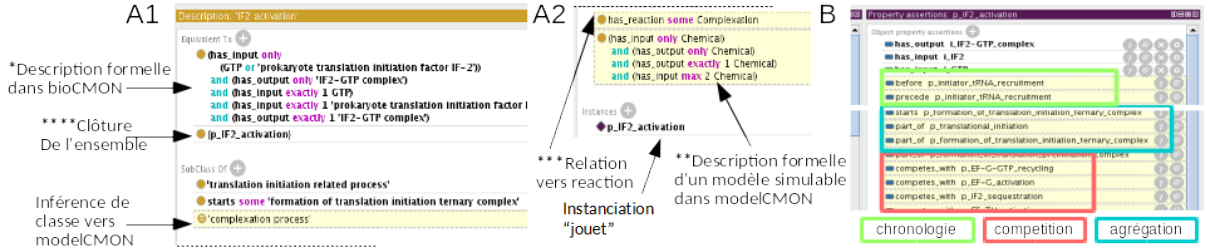


FIGURE 3 – Capture d'écran de Protégé explorant l'activation de l'IF2 (A : classe, B : instance).

2.3 Enrichissement de la connaissance de bioCMON par inférence

Les classes de C'MON ont été conçues dans le cadre d'une expressivité en logique de description SROIQ et certaines inférences au niveau des concepts peuvent être de grande complexité. Pour apporter automatiquement des informations complémentaires à notre modèle, chacune des classes feuilles de bioCMON a été instanciée par des éléments "jouets" et les ensembles ont été clôturés aux seules instances de ces classes (figure 3A****). Dans un second temps, profitant de la formalisation de C'MON en OWL2 (Cuenca Grau *et al.* 2008), des règles en *Semantic Web Rule Language* (SWRL) ont été conçues (Krisnadhi *et al.* 2011). Par exemple :

```
macromolecule(?x), input_of(?x,?p1), input_of(?x,?p2), DifferentFrom (?p1,?p2)
    -> competes_with(?p1,?p2)
```

Ces règles formalisent de nouvelles relations qui permettent de faire ressortir des informations de régulation de processus au travers de compétitions (*compets_with*) ou de chronologies relatives (*before* ou *precedes*) entre des processus et enfin de filtrer les entités biologiques participantes en fonction de l'agrégation des processus (*has_filtered_input*) profitant ainsi pleinement de l'approche systémique multi-échelle (figure 3B).

3 Conclusion et perspectives

C'MON a pour objectif une vision complète à différentes granularités des processus cellulaires. Actuellement C'MON comprend 1384 concepts qui permettent la modélisation des

processus impliqués dans l'expression génique (incluant les processus de régulation). Cette organisation se structure le long d'une échelle de processus de haut niveau (*e.g.* : traduction) aux processus les plus élémentaires (*e.g.* : activation de l'IF2) reliés par des sous-propriétés de *part_of*. Les processus, entités biologiques participantes, modèles et paramètres associés ont été formellement décrits et reliés entre eux. Chacun des processus est relié à un type de réaction et des paramètres [comme un type de constante d'affinité obtenue expérimentalement (k-on) ou estimée à partir de données biologiques (k-est) ; Figure 2]. De plus, C'MON s'enrichit automatiquement d'informations biologiques fonctionnelles complémentaires (chronologie, agrégation, compétition) suite à une inférence sur les instances (figure 3B). En plus d'un effort de curation, ce travail s'appuie sur l'intégration d'ontologies préexistantes (GO, ChEBI, KEGG, SO et SBO). Dans cette démarche, la conservation des IRI et Id des concepts intégrés vise à proposer un modèle interopérable. Ainsi l'utilisation de C'MON peut permettre la réexploitation de travaux préexistants sur l'expression génique en microbiologie avec les nouveaux apports de notre modèle.

Le modèle ontologique C'MON décrit dans cette article montre donc que l'approche systémique multi-échelle est bien pertinente pour décrire les processus cellulaires et les relier à leurs paramètres. C'MON ouvre ainsi de nouvelles perspectives pour l'intégration des données biologiques hétérogènes dans l'objectif d'une utilisation interdisciplinaire.

La preuve de concept étant présentée ici, les prochaines étapes consistent à compléter C'MON avec l'ensemble des processus cellulaires bactériens et leur régulation, comme ceux de la réplication de l'ADN, puis de l'étendre aux processus cellulaires eucaryotes.

A terme, C'MON sera associée à un entrepôt de données biologiques omiques dans le cadre d'un système d'information.

Remerciement : Ce travail a été financé par le projet LIDEX IMSV de Paris-Saclay.

Références

ALTMAN T. *et al.* (2013). A systematic comparison of the MetaCyc and KEGG pathway databases. BMC Bioinformatics 14 :112.

COURTOT M *et al.* (2011). Controlled vocabularies and semantics in systems biology. Molecular Systems Biology 7 :543.

CUENCA GRAU B. *et al.* (2008). OWL 2: The Next Step for OWL. Journal of Web Semantics 6.

DE MATOS P *et al.* (2010). Chemical Entities of Biological Interest: an update. Nucleic Acids Res 38, p. 249–254.

EILBECK K. *et al.* (2005). The Sequence Ontology: A tool for the unification of genome annotations. Genome Biol. 6:R44 .

GLIMM B. *et al.* (2014). HermiT : An OWL 2 Reasoner. J. of Automated Reasoning 53, p. 245–269.

GOELZER A. & FROMION V. (2011). Bacterial growth rate reflects a bottleneck in resource allocation. Biochim Biophys Acta. 1810, p. 978-988.

GOELZER A. *et al.* (2015). Quantitative prediction of genome-wide resource allocation in bacteria. Metab Eng. 32, p. 232-243.

HILL *et al.* (2013). Dovetailing biology and chemistry: integrating the Gene Ontology with the ChEBI chemical ontology. BMC Genomics 14 :513.

KITANO H. *et al.* (2001). Foundations of Systems Biology. MIT Press.

KRISNADHI A. *et al.* (2011). OWL and Rules. Reasoning Web. 6848, p. 382-415.

METZKER M.L. (2010) Sequencing technologies-the next generation. Nat. Rev. Genet. 11, p. 31–46.

NEKRUTENKO A. &, TAYLOR J. (2012). Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat. Rev. Genet. 13, p. 667–672.

THE GENE ONTOLOGY CONSORTIUM (2000). Gene ontology: tool for the unification of biology. Nat Genet 25, p.25–29.

Anti-patrons partiels pour l'identification des problèmes de contradiction sociale dans une ontologie

Mounira Harzallah

LINA, Université de Nantes
mounira.harzallah@univ-nantes.fr

Résumé : Dans le contexte du web sémantique et des masses de données, les ontologies nécessaires à faire communiquer les objets du web seront probablement de grande taille et construites à partir des ressources diverses et hétérogènes. Leur validation sera plus que jamais primordiale. Dans cet article, nous mettons en exergue les liens entre deux types problèmes pouvant nuire à la qualité d'une ontologie : l'insatisfiabilité d'artéfacts et le problème de contradiction sociale. Nous proposons des anti-patrons partiels et une heuristique pour aider à identifier les problèmes de contradiction sociale, tout en minimisant l'intervention humaine.

Mots clés : Validation d'ontologie, Insatisfiabilité, Problème de contradiction sociale, Anti-patron partiel.

1 Introduction

Dans le contexte du web sémantique et des masses de données reliées sur internet, les ontologies nécessaires à faire communiquer les objets du web seront probablement de grande taille et construites à partir des ressources diverses et hétérogènes. Leur validation sera plus que jamais primordiale et complexe. Dans nos travaux, nous considérons la validation comme un processus qui devrait être réalisé en parallèle avec celui d'acquisition/extraction des artefacts de l'ontologie (Harzallah et al. 2015). Il devrait être réalisé le plus tôt possible afin d'éviter la propagation des problèmes dans une ontologie et rendre complexe sa correction. En outre, nous considérons la validation comme un processus : 1) qui cherche une mauvaise qualité dans une ontologie et 2) qui propose ensuite une méthode pour améliorer cette qualité en identifiant et en enlevant les causes liées à cette mauvaise qualité. Des méthodes ont été proposées pour l'évaluation de la qualité d'une ontologie. Actuellement, il n'y a pas encore un consensus sur comment une ontologie doit être évaluée (Neuthaus & Vizedom, 2013). En plus, certains problèmes qui peuvent nuire à la qualité d'une ontologie sont peu traités et ne peuvent être identifiés que par un acteur social. Nous nous intéressons aux problèmes de contradiction sociale (i.e. une contradiction entre la perception qu'un acteur social a d'une ontologie et le contenu de cette ontologie), par exemple, dans une ontologie une relation de subsumption jugée fausse par un acteur social. Nous cherchons à identifier ces problèmes en les reliant aux problèmes d'insatisfiabilité qui sont identifiables d'une façon automatique.

Dans cet article, nous discutons des méthodes d'identification des problèmes pouvant affecter la qualité d'une ontologie. Ensuite, nous proposons des anti-patrons partiels de l'insatisfiabilité et une heuristique pour détecter des problèmes de contradiction sociale.

2 Méthodes d'identification des problèmes affectant la qualité d'une ontologie

Plusieurs termes sont utilisés pour évoquer la notion de problème affectant la qualité d'une ontologie : 1) les erreurs de taxonomie (Gomez-Perez et al. 2001) ou les erreurs structurelles (Buhmann et al. 2011), 2) les anomalies de conception ou défaillances (Baumeister & Seipel,

2005), 3) les anti-patrons (Roussey et al. 2010), (Buhmann et al. 2011), 4) les embûches ou les mauvaises pratiques (Poveda et al. 2012) et 5) les défauts logiques (Buhmann et al. 2011). Vu l'hétérogénéité et la diversité de ces termes et le recouvrement de leur définition, nous avons proposé une typologie des problèmes permettant de standardiser leur définition et d'aider à les identifier (Gherasim, 2013). Cette typologie est définie selon deux dimensions : aspect logique vs aspect social de l'ontologie et erreurs vs situations indésirables. Les erreurs sont des problèmes qui rendent une ontologie inutilisable et les situations indésirables sont ceux qui l'affectent sans empêcher son utilisation. Les problèmes logiques sont reliés à l'interprétation que donnent les machines à une ontologie sur la base de ses spécifications logiques. Les problèmes sociaux sont reliés à la perception/interprétation qu'un acteur social a d'une ontologie. Le terme « acteur social », emprunté à la terminologie utilisée dans le modèle de qualité des modèles conceptuels SEQUAL (Krogstie et al. 1995), mentionne une partie prenante (un individu ou un groupe d'individus) concernée par l'ontologie en question.

Nous avons formalisé la majorité des problèmes logiques en logiques de description en considérant les notions synthétisées dans Guarino et al. (2009) i.e. Interprétation (I), Modèles intentionnels (Intended Models) (IM), Langage (L), Ontologie (O), et les 2 relations : $\models$ et $\vdash$ (Gherasim, 2013). Nous formalisons certains problèmes sociaux, en rajoutant la notion de SAP_O (Social Actor Perception of Ontology) par analogie avec les notions de IM. Par exemple, $SAP_O \not\models \varphi$ implique φ est fausse dans la perception d'un acteur social de O.

Plusieurs travaux se sont intéressés à l'étude des problèmes pouvant affecter une ontologie et les techniques de leur identification. Des raisonneurs (Pellet, FactC++, Racer) ont été développés pour identifier certains problèmes logiques, particulièrement les problèmes d'inconsistance, d'insatisfiabilité ou de redondance, mais sans déterminer leur origine. Des travaux relativement récents ont cherché à identifier et à présenter à l'utilisateur le plus petit ensemble d'axiomes menant à une insatisfiabilité (Wang et al. 2005) (Rodler et al. 2013). Pour déterminer la cause ayant la meilleure chance d'être la bonne et en minimisant le nombre de questions à poser à un utilisateur, Rodler et al. (2013) utilisent une méthode probabiliste prenant en compte les causes les plus fréquentes. Des nouvelles heuristiques cherchent à détecter des problèmes qui ne sont pas nécessairement détectés par des raisonneurs, par exemple des candidates à l'absence d'exclusivité (disjointness) (Quazi & Quadir, 2011). Roussey et al. (2010) ont défini des anti-patrons pour l'identification de certains problèmes d'insatisfiabilité ou des candidats à des problèmes d'ontologie inadaptée ou d'ontologie non minimale. L'identification d'autres problèmes logiques, par exemple l'incomplétude d'un point de vue logique, demande la génération de IM, ce qui n'est pas toujours évident à avoir. Concernant les problèmes sociaux, des approches récentes utilisent des systèmes de question/réponse pour les identifier (Pammer, 2010). Les questions portent souvent sur la totalité des artefacts d'une ontologie ce qui permet une validation complète mais qui repose sur l'utilisateur qui pourrait commettre des erreurs en répondant à des centaines de questions. Poveda et al. (2012) ont développé l'outil OOPS ! pour aider à détecter certains de ces problèmes, mais il n'est pas possible de comprendre clairement ce que cet outil peut détecter.

Dans la suite, nous présentons notre approche pour la définition d'anti-patrons partiels pour les problèmes d'insatisfiabilité et de contradiction sociale. Nos travaux complètent ceux portant sur l'identification des cas susceptibles de correspondre à des problèmes, par exemple, les travaux de Roussey et al. (2010) ou de Quazi & Qadir (2011) et ceux qui font intervenir l'utilisateur pour l'aide à l'identification des problèmes.

3 Formalisation des problèmes de contradiction sociale et d'insatisfiabilité

Dans une ontologie, certains problèmes peuvent être dus à ou impliquer l'existence d'un autre problème. L'identification des liens entre problèmes est un moyen pour bien optimiser le processus de validation d'une ontologie (Harzallah et al. 2015). Par exemple : un problème de redondance de relations de subsumption peut être dû à l'existence d'une relation de subsumption fausse, qui pourrait correspondre à un problème de contradiction sociale, et

l'analyse de cette redondance peut aider à cibler cette relation. Nous avons cherché à déterminer des liens entre des problèmes en se basant sur leur formalisation. En effet, il s'agit pour deux problèmes P1 et P2 et leur formule respective F(P1) et F(P2) de déterminer quelle information F nécessaire pour que F(P1) & F soit équivalente à F(P2). Des liens existent d'une façon intuitive entre les problèmes d'insatisfiabilité et les problèmes d'inconsistance logique, d'ontologie inadaptée ou de contradiction sociale. En effet, les problèmes d'insatifiabilité représentent des artefacts dont l'interprétation est vide. Cependant, on définit rarement ce type d'artefacts. Ceci implique que probablement, on possède des instances pour l'artefact insatisfiable et en les rajoutant à l'ontologie on produit un problème d'inconsistance. Ce dernier cache souvent l'existence de quelque chose qui est fausse par rapport à des IM (problème d'ontologie inadaptée) ou par rapport à la perception d'un acteur social (problème de contradiction sociale). Par exemple, le problème d'insatisfiablité P36 (Table 1) dans une ontologie O où F(P36)= **{**$O \models Ci \sqsubseteq \forall R.Cj$, $Ci \sqsubseteq \exists R.C_k$, $C_j \sqcap C_k \equiv \bot$} peut impliquer l'existence :

- des problèmes d'inconsistance logique. En effet, le rajout de l'information $\{x \in Ci^I\}$ dans cette ontologie engendre une inconsistance (Ci^I étant une interprétation de Ci) ;
- des problèmes d'ontologie inadaptée. En effet, la possession de l'information : $IM \models Ci \sqsubseteq \exists R.\neg Cj$, $IM \models Ci \sqsubseteq \forall R.\neg C_k$ ou $IM \models not(C_j \sqcap C_k \equiv \bot)$ (IM étant un modèle intentionnel de L) implique l'existence d'un cas d'ontologie inadaptée dans O ;
- des problèmes de contradiction sociale. En effet, la possession de l'information : $SAP_O \models not\ (Ci \sqsubseteq \forall R.Cj)$, $SAP_O \models not\ (Ci \sqsubseteq \exists R.C_k)$ ou $SAP_O \models not\ (C_j \sqcap C_k \equiv \bot)$ implique l'existence d'un cas de contradiction sociale dans O.

Pour formaliser la relation de « un problème peut impliquer l'existence d'un autre problème » entre les problèmes d'insatisfiabilité et de contradiction sociale, nous avons, tout d'abord, utilisé notre typologie pour classifier 60 cas de problèmes traités dans la littérature (nous les avons numérotés de P1 à P60) : 33 cas (de P1 à P33) traités dans le catalogue de Poveda et al. (2012), 11 cas (de P34 à P44) considérés dans (Roussey et al. 2010), 10 cas (de P45 à P54) proposés dans (Wang et al. 2005) et présentés dans (Buhmann et al. 2011), 6 cas (de P55 à P60) proposés dans (Fahad et al. 2008). Cette classification a montré que la majorité de ces cas correspond à un des 5 types de problèmes suivants : Insatisfiabilité (12 cas), Non minimale (12 cas), Contradiction sociale (11 cas), Erreurs de conception (12 cas), et Incomplétude du point de vue social (9 cas). Nous nous sommes focalisés sur les cas d'insatisfiabilité et de contradiction sociale. Parmi les 12 cas d'insatisfiabilité étudiés, seulement 8 cas sont différents (lignes grisées dans la table 1).

TABLE 1 - *Problèmes d'insatifiabilité traités dans la littérature*

Problèmes d'insatisfaisabilité	Formalisation
P34-Anti-pattern AndisOR	$O \models C \sqsubseteq \exists R.(C_i \sqcap C_j)$, $C_i \sqcap C_j \equiv \bot$
P36 - Anti-pattern UniversalExistence	$O \models C \sqsubseteq \forall R.C_i$, $C \sqsubseteq \exists R.C_j$, $C_i \sqcap C_j \equiv \bot$
P37 - Anti-pattern EquivalenceIsDifference	$O \models C_i \equiv C_j$, $C_i \sqcap C_j \equiv \bot$
P45 - Partition error	$O \models C \sqsubseteq (C_i \sqcap C_j)$, $C_i \sqcap C_j \equiv \bot$
P47 - Having both a class and its complement as super condition	$O \models C \sqsubseteq (C_i \sqcap \neg C_i)$
P48 - Having a super condition that is assumed to be disjoint	$O \models C \sqsubseteq \neg \top$
P49 - Having a super condition that is an existential restriction that has a filler which is disjoint with the range of the restricted property	$O \models C \sqsubseteq \exists R.C_i$, $R.Range \equiv C_j$, $C_j \sqcap C_i \equiv \bot$
P50 - Having an universal restriction with Nothing as the filler and a must existing restriction along property relationships.	$O \models C \sqsubseteq \forall R.\bot$, $C \sqsubseteq \exists R.C_i$, $C_i \neq \bot$
P51 - Having more than allowed existential restrictions	$O \models C \sqsubseteq <2R.\top$, $C \subseteq \exists R.C_i$, $C \sqsubseteq \exists R.C_j$, $C_i \sqcap C_j \equiv \bot$
P52 - Having a super condition containing conflicting cardinality restrictions	$O \models C \sqsubseteq > nR.\top$, $C \subseteq <nR.\top$
P53 - Inconsistence from other ressources	$O \models C \sqsubseteq \exists R.C_i$, $C_i \sqsubseteq C_j$, $C_j \equiv \bot$
P54 - Having a super condition that is an existential restriction where the domain of the restricted property is disjoint with it	$O \models C_i \sqsubseteq \exists R.T$, $R.Domain \equiv C_j$, $C_i \sqcap C_j \equiv \bot$

La table 2 comprend la formalisation des 11 cas de problème de contradiction sociale. Leur analyse a mis en évidence l'existence de 6 cas d'égalité (P05, P19, P25, P27, P31, P41) et 3 cas d'inclusion d'artefacts (P14, P15, P46) qu'un acteur social juge faux. Dans 4 cas (P14, P15, P19, P41) sur 11, une correction probable du problème a été proposée dans la littérature.

TABLE 2 - *Problèmes de contradiction sociale*

Cas de problème de contradiction sociale	Formalisation
P05 - Wrong inverse relationship	$O \models R^{-1}{}_i \equiv R_i$, $SAP_O \not\models R^{-1}{}_i \equiv R_i$
P14 - Misusing "Owl :allvaluesFrom"	$O \models C \sqsubseteq \forall R.C_i$, $SAP_O \not\models C \sqsubseteq \forall R.Ci$, $SAP_O \models C \sqsubseteq \exists R.C_j$,
P15 - Misusing "not some" and "some not"	$O \models C \sqsubseteq \neg(\exists R.C_j)$, $SAP_O \not\models C \sqsubseteq \neg\exists R.C_j$, $SAP_O \models C \sqsubseteq \exists R.\neg C_j$
P19 - Swapping intersection and union	$O \models R.range \equiv (C_i \sqcap C_k)$, $SAP_O \not\models R.range \equiv (C_i \sqcap C_k)$, $SAP_O \models R.range \equiv (C_i \sqcup C_k)$
P25 - Define a relationship inverse to itself	$O \models R^{-1} \equiv R$, $SAP_O \not\models R^{-1} \equiv R$
P27 - Defining wrong equivalent relationship	$O \models R_i \equiv R_i$, $SAP_O \not\models R_i \equiv R_i$
P28 - Defining wrong symmetric relationship	$O \models$ Symmetric(R) , $SAP_O \not\models$ Symmetric(R)
P29 - Defining wrong transitive relationship	$O \models$ Transitive(R) , $SAP_O \not\models$ Transitive(R)
P31 - Defining wrong equivalent classes	$O \models C_i \equiv C_i$ $SAP_O \not\models C_i \equiv C_i$
P41 - DisjointnessOfComplement	$O \models C_j \equiv \neg C_i$, $SAP_O \not\models C_i \equiv \neg C_i$, $SAP_O \models C_i \sqcap C_j \equiv \bot$
P46 - Semantic inconsistency	$O \models C_i \sqsubseteq R.C_i$, $SAP_O \not\models C_i \sqsubseteq R.C_i$

4. Méthode de détection des candidats à l'insatisfiabilité et à la contradiction sociale

En se basant sur le principe qu'un problème d'insatisfiabilité « peut impliquer l'existence de » un problème de contradiction sociale, nous avons extrait, à partir de la formalisation de chacun des 8 cas différents d'insatisfiabilité, les axiomes pouvant représenter chacun une contradiction sociale (i.e. un axiome pouvant être faux selon un acteur social). Par exemple, P36 où $F(P36)=\{C \sqsubseteq \forall R.Ci , C \sqsubseteq \exists R.C_j , C_i \sqcap C_j \equiv \bot\}$, peut impliquer l'existence d'une ou de plusieurs contradictions sociales parmi les suivantes : ($SAP_O \not\models C \sqsubseteq \forall R.Ci$), ($SAP_O \not\models C \sqsubseteq \exists R.C_j$) ou ($SAPo \not\models C_i \sqcap C_j \equiv \bot$). Neuf axiomes types ont été extraits (on note cet ensemble A_{Ins}). Trois axiomes parmi eux représentent des problèmes de contradiction sociale connus (P14, P15, P31). Nous avons ordonné ces axiomes selon l'ordre décroissant de notre estimation de la probabilité qu'un axiome corresponde à une contradiction sociale. Cette estimation dépend de la difficulté de la compréhension d'un axiome (critère évoqué dans (Roussey et al. 2010)) et de la pertinence de sa considération dans un processus de validation Ce deuxième critère a été jugé en fonction de la rareté d'un axiome dans une ontologie et de son appartenance à plusieurs cas d'insatisfiabilité. L'ordre obtenu (noté Ord_{Ains}) est le suivant : A1 : $C_i \sqcap C_j \equiv \bot$, A2 : $C_i \sqsubseteq \exists R.C_j$, A3 : $C_i \sqsubseteq \forall R.C_j$, A4 : $C_i \sqsubseteq <nR.C_j$, $n>1$, A5 : $C_i \sqsubseteq >nR.C_j$, $n>1$, A6 : $C_i \equiv C_j$, A7 : $R.Domain \equiv C_j$, A8 : $R.Range \equiv C_i$, A9 : $C_i \sqsubseteq C_j$.

A partir des 8 cas différents d'insatisfiabilité, nous avons extrait également 13 anti-patrons partiels de l'insatisfiabilité (APPI). Un APPI est un anti-patron de l'instatifiabilité auquel on a enlevé un axiome (son axiome manquant). Il aide à identifier une insatisfiabilité et une contradiction sociale : si on est proche d'une insatisfiabilité on est proche d'une contradiction sociale. Un APPI est à utiliser quand les problèmes d'insatisfiabilité sont déjà corrigés dans une ontologie O. En plus, il n'est intéressant que s'il n'y a rien qui contredit son axiome manquant dans O. Dans ce cas, pour engendrer une insatifiabilité, on complète l'occurrence de chaque APPI (OAPPI) dans O, par l'axiome manquant si la validité de cet axiome est confirmée par un acteur social. Pour détecter une contradiction sociale, on vérifie la validité des axiomes qui composent chaque OAPPI. Dans les deux cas, l'APPI « $C_i \sqsubseteq C_j$ » n'est pas à prendre en compte car il est très fréquent dans une ontologie.

Nous proposons l'heuristique HPCS_APPI (Heuristique pour l'identification des Problèmes de Contradiction Sociale à l'aide des APPI) qui utilise les axiomes de A_{Ins} afin d'aider à cibler des cas de contradiction sociale, tout en cherchant à minimiser l'intervention humaine. Il s'applique à une ontologie ne comprenant pas des problèmes d'insatisfabilité, sinon il faut tout d'abord les corriger, et il cherche des contradictions sociales dans chaque OAPPI, sans la transformer en une insatisfiabilité.

TABLE 3 - *Anti-patrons partiels de l'insatisfiabilité*

Anti-patron partiel à 3 axiomes	Axiome manquant
$C_i \sqcap C_j \equiv \bot$, $C \sqsubseteq \exists R.C_i$, $C \sqsubseteq <2R.\top$	$C \sqsubseteq \exists R.C_i$
$C_i \sqcap C_j \equiv \bot$, $C \sqsubseteq \exists R.C_i$, $C \sqsubseteq \exists R.C_i$	$C \sqsubseteq <2R.\top$
$C \subseteq \exists R.C_i$, $C \sqsubseteq \exists R.C_i$, $C \sqsubseteq <2R.\top$	$C_i \sqcap C_i \equiv \bot$
$C_i \sqcap C_i \equiv \bot$, $C \sqsubseteq \forall R.C_i$,	$C \sqsubseteq \exists R.C_i$
$C_i \sqcap C_i \equiv \bot$, $C \sqsubseteq \exists R.C_i$	$C \sqsubseteq \forall R.C_i$
$C \sqsubseteq \exists R.C_i$, $C \sqsubseteq \forall R.C_i$	$C_i \sqcap C_i \equiv \bot$
$C_i \sqsubseteq \exists R.T$, $R.Domain \equiv C_i$	$C_i \sqcap C_i \equiv \bot$
$C_i \sqcap C_i \equiv \bot$, $R.Domain \equiv C_i$	$C_i \sqsubseteq \exists R.T$
$C_i \sqcap C_i \equiv \bot$, $C_i \subseteq \exists R.T$	$R.Domain \equiv C_j$
$C \sqsubseteq \exists R.C_i$, $R.range \equiv C_i$	$C_i \sqcap C_i \equiv \bot$,
$C_i \sqcap C_i \equiv \bot$, $R.range \equiv C_i$	$C \sqsubseteq \exists R.C_i$
$C_i \sqcap C_i \equiv \bot$, $C \sqsubseteq C_i$	$C \sqsubseteq C_j$
$C \sqsubseteq C_i$, $C \sqsubseteq C_j$	$C_i \sqcap C_i \equiv \bot$,
$C_i \sqcap C_i \equiv \bot$	$C_i \equiv C_j$ ou $C_i \sqsubseteq C_j$
$C_i \equiv C_j$	$C_i \sqcap C_i \equiv \bot$
$C \sqsubseteq \exists R.(C_i \sqcap C_i)$	$C_i \sqcap C_i \equiv \bot$
$C_i \sqsubseteq C_j$	$C_i \sqcap C_i \equiv \bot$

Soit O une ontologie, HPCS_APPI cherche dans O les OAPPI et les classe en trois ensembles Fm (m de 1 à 3), chacun comprend des OAPPI composées du même nombre d'axiomes. Soit n_m le nombre de OAPPI dans F_m. Chaque $OAPPI_{mi}$ (i de 1 à n_m) correspond à une OAPPI avec m axiomes. Les axiomes A_{ij} de chaque $OAPPI_{mi}$ (j de 1 à m) sont ordonnés selon l'ordre Ord_{Ains} défini précédemment, tel que si A_{ij}, et A_{ik} deux axiomes de $OAPPI_{mi}$ et $j<k$ alors A_{ij} appartient à un type d'axiome d'indice inférieur ou égal à celui du type d'axiome auquel appartient A_{ik}. Dans cette heuristique, on traite les OAPPI dans l'ordre décroissant du nombre d'axiomes qui les composent (plus le nombre d'axiomes d'une OAPPI est grand plus la chance de détecter une contradiction dans cette OAPPI est importante) et les axiomes de chaque OAPPI sont vérifiés dans l'ordre Ord_{Ains}. Si un axiome d'une OAPPI est vrai, on indique dans O qu'il est validé et on continue à vérifier les autres axiomes de la même occurrence. S'il est faux, on indique qu'il est faux dans O et on enlève toutes les OAPPI à lesquelles il appartient. HPCS_APP est définie comme suit :

```
Début HPCS_APPI
 Pour m= 3 à 1 FAIRE, Pour i=1 à nm Faire
    Vérifier que OAPPImi appartient à Fm, si oui
    Pour j=1 à m
       si Aij est validé alors aller à Fin pour j
       sinon vérifier avec un acteur social si Aij est Vrai
            Si Aij est faux d'après un acteur social alors
               supprimer Aij de O et rajouter (not Aij) à O
               si Aij appartient à un des OAPPI alors F=F/OAPPI
               aller à Fin pour i
            Sinon noter que Aij est validé dans tous les OAPPI de F
    Fin pour j
   Fin pour i, Fin pour m
Fin HPCS_APPI
```

5. Conclusion

Plusieurs travaux se sont intéressés aux problèmes logiques qui nuisent à la qualité d'une ontologie. Les problèmes qui ne peuvent être détectés que par un acteur social sont peu considérés dans la littérature. Dans cet article, nous avons proposé des anti-patrons partiels et

une heuristique pour aider à bien cibler des candidats aux problèmes de contradiction sociale tout en minimisant l'intervention humaine. Actuellement, nous sommes en train d'expérimenter notre approche. Nous cherchons, d'une part, à valider l'intérêt de l'ordre de vérification des axiomes de A_{ins} et celui de traitement des OAPPI pour la minimisation de l'intervention humaine et d'autre part à évaluer le nombre de questions à poser à un acteur social par rapport aux nombres des problèmes corrigés et des problèmes ignorés.

Comme perspectives, nous allons intégrer les APPI et notre heuristique dans notre approche globale de validation d'ontologie (Harzallah et al. 2015), en étudiant à quel moment opportun, il faut identifier et corriger les problèmes de contradiction sociale afin d'améliorer l'efficience du processus de validation d'ontologie dans sa globalité

Références

BAUMEISTER J., SEIPEL, D. (2005). Smelly owls-design anomalies in ontologies. pp 215–220 of: Proc. of 18th int. florida artificial intelligence research society conf.

BUHMANN L., DANIELCZYK, S. & LEHMANN, J. (2011). D3.4.1 report on relevant automatically detectable modelling errors and problems. Tech. rept. LOD2-Creating Knowledge out of Interlinked Data.

FAHAD M., M.ABDUL QADIR & S. A. H. SHAH (2008). Evaluation of ontologies and dl reasoners. In Zhongzhi Shi, E. Mercier-Laurent, and D. Leake, editors, Intelligent Information Processing IV, volume 288 of IFIP : The International Federation for Information Processing, pp 17–27.

GHÉRASIM T.D. (2013). Détection de problèmes de qualité dans des ontologies construites automatiquement à partir de textes. Thèse en Informatique de l'université de Nantes.

GOMEZ-PEREZ A., FERNANDEZ-LOPEZ, M., CORCHO, O. (2001). Ontological engineering:With examples from the areas of knowledge management, e-commerce and the semantic web. Adv. Inf. And Know. Processing. Springer.

GUARINO N., OBERLE, D., & STAAB, S. (2009). What is an ontology? pp 1–17 of: Studer, R., & Staab, S. (eds), Handbook on ontologies, 2 edn. International Handbooks on Information Systems. Springer.

HARZALLAH M., G. BERIO, & KUNTZ P. (2015). Towards an Approach for Configuring Ontology Validation. In Knowledge Discovery, Knowledge Engineering and Knowledge Management Vo.553 of the series Communications in Computer and Information Science. pp 388-404, Springer.

KROGSTIE, J., O. LINDLAND & G. SINDRE (1995). Defining quality aspects for conceptual models. In Proceedings of the IFIP8.1 Working Conference on Information Systems Concepts : Towards a Consolidation of Views (ISCO3), pp 216–231.

NEUTHAUS F. & A VIZEDOM. (2013). Towards Ontology Evaluation across the Life Cycle. In ontolog.cim3.net/file/work/OntologySummit2013/OntologySummit2013,_Communiqué.

PAMMER V. (2010). Automatic Support for Ontology Evaluation Review of Entailed Statements and Assertional Effects for OWL Ontologies. PhD thesis, Graz University of Technology.

POVEDA M., SUAREZ-FIGUEROA, M.C. & GOMEZ-PEREZ, A. (2012). Validating ontologies with oops !, dans Proceedings of the 18th international conference on Knowledge Engineering and Knowledge Management (EKAW'12), Springer-Verlag, pp. 267– 281.

QAZI N. I. & M. A. QADIR (2011). Algorithms for the evaluation of ontologies for extended error taxonomy and their application on large ontologies. J. UCS, 17(7) pp:1005–1020.

RODLER P., K. SHCHEKOTYKHIN, P. FLEISS & FRIEDRICH G. (2013). Rio: Minimizing user interaction in ontology debugging. In Wolfgang Faber and Domenico Lembo, editors, Web Reasoning and Rule Systems, volume 7994 of Lecture Notes in Computer Science, pp 153–167. Springer Berlin Heidelberg.

ROUSSEY C., F. SCHARFFE, O. CORCHO, AND O. ZAMAZAL. (2010). Une méthode de débogage d'ontologies OWL basées sur la détection d'anti-patrons. In Sylvie Desprès, editor, IC2010, 21éme Journées francophones d'Ingénierie des Connaissances, Collection mathématiques & Informatique, pp 43–54. Presses de Nimes.

WANG H., M. HORRIDGE, A. RECTOR, N. DRUMMOND & J SEIDENBERG (2005). Debugging owl-dl ontologies: A heuristic approach. In Y Gil, E Motta, V.R Benjamins, and M.A. Musen, editors, The Semantic Web, ISWC 2005, volume 3729 of Lecture Notes in Computer Science, pp 745–757. Springer Berlin Heidelberg.

Ingénierie des connaissances et texte

Exploiter la structure discursive du texte pour valider les relations candidates d'hyperonymie issues de structures énumératives parallèles

Mouna Kamel, Cassia Trojahn

IRIT UMR 5505 Institut de Recherche en Informatique de Toulouse, Toulouse, France
{mouna.kamel, cassia.trojahn}@irit.fr

Résumé : L'acquisition automatique de connaissances à partir de textes est un enjeu majeur pour la construction de ressources sémantiques. L'une des tâches cruciales concerne l'identification des relations sémantiques. Cette tâche peut être déclinée en trois phases : l'identification de relations dites candidates, la validation de ces relations et l'intégration de ces relations au sein d'une ressource existante ou en cours de construction. Nous intéressons ici à la validation automatique de relations candidates extraites de structures textuelles spécifiques, les structures énumératives parallèles, en tirant profit de leurs propriétés discursives. L'approche proposée repose sur l'exploitation combinée d'un réseau sémantico-lexicale et une ressource distributionnelle. Les résultats obtenus montrent une exactitude comprise entre 0.5 et 0.67 selon les conditions expérimentales.

Mots-clés : relations sémantiques, validation, structures discursives, réseau lexico-sémantique, ressource distributionnelle

1 Introduction

Le processus d'extraction de relations sémantiques à partir de texte est une tâche cruciale car en amont des processus de construction de ressources sémantiques. Ce processus se fait généralement en trois étapes : repérer les relations candidates dans le texte, valider ces relations, et, pour celles qui ont été validées, les intégrer dans une ressource existante ou en cours de construction. La première étape a fait l'objet de nombreux travaux et de nombreuses approches ont été proposées (approches linguistiques, statistiques, mixtes, avec ou sans apprentissage). Des erreurs peuvent cependant se produire, dues à la stratégie mise en œuvre (patrons lexico-syntaxiques insuffisamment contraints, exactitude des techniques d'apprentissage inférieure à 100% , etc.), ou encore aux différents outils de traitement automatique des langues (ou TAL) appliqués successivement dans les phases de pré-traitement. Aussi, la deuxième étape relative à la validation des relations précédemment identifiées s'avère indispensable. La troisième étape consiste à intégrer les relations validées au sein d'une ressource existante, en s'appuyant, sur des approches d'alignement terminologiques, structurelles ou sémantiques.

Dans cet article nous nous intéressons à la tâche de validation de relations candidates, qui consiste à confirmer ou non ces relations candidates. Plusieurs approches existent, des approches manuelles qui font appel à une expertise humaine, et des approches automatiques qui consistent soit à s'appuyer sur des ressources sémantiques externes, soit à procéder par renforcement si l'on dispose de gros corpus, à l'échelle du web. L'approche que nous proposons s'inscrit dans la continuité de travaux visant à extraire les relations sémantiques, en l'occurrence les relations d'hyperonymie, portées par les structures énumératives parallèles (appelées par la

suite SEP) (Fauconnier & Kamel, 2015). Le choix de s'intéresser aux SEP est motivé par les raisons suivantes : (1) elles sont fréquentes en corpus, notamment dans les textes scientifiques ou encyclopédiques qui sont des textes appropriés pour la construction de ressources ; dans ce type de texte, les SEP sont très souvent exprimées à l'aide de moyens de mise en forme (caractères typographiques et dispositionnels) pour faciliter l'effort cognitif du lecteur - ces SEP sont alors hors de portée des outils classiques de TAL, (2) elles sont souvent porteuses de relations hiérarchiques, relations qui constituent l'ossature des ressources sémantiques (Buitelaar *et al.*, 2005), (3) elles possèdent des propriétés discursives bien établies qui leur confèrent une unité sémantique (Virbel, 1989; Pascual, 1991).

L'approche que nous présentons ici, bien qu'elle s'appuie aussi sur des ressources externes, est originale dans la mesure où elle exploite les propriétés discursives de la SEP, à l'aide de deux types de ressources externes complémentaires, un réseau lexico-sémantique et une ressource distributionnelle. Le réseau lexico-sémantique permet alors de valider les relations spécifiées dans ce réseau, alors que la ressource distributionnelle favorise la spécification de nouvelles relations. De plus, l'unité sémantique dont bénéficie la SEP permet de désambiguïser les relations.

Outre le fait que l'approche que nous proposons a pour objet de valider des relations, elle peut également être utilisée pour valider les systèmes d'extraction de relations à partir de SEP. Par ailleurs, les nouvelles relations validées à l'aide de la ressource distributionnelle constituent alors une source d'enrichissement pour le réseau lexico-sémantique. Enfin, bien qu'implémentée et évaluée pour la langue française, elle reste reproductible pour toute autre langue.

L'article est organisé de la façon suivante. La section 2 fait état des méthodes de validation de relations généralement utilisées. La section 3 définit la SEP, énonce ses propriétés, et donne une représentation de son schéma discursif selon la Rhetorical Structure Theory. La section 4 présente les principes sur lesquels se base l'approche adoptée, ainsi que les algorithmes mis en œuvre. L'application et les résultats de l'évaluation sont décrits dans la section 5, suivis d'une discussion. Nous concluons et proposons quelques perspectives à ce travail en section 6.

2 Etat de l'art

Valider une relation identifiée en corpus consiste à confirmer le sens véhiculé par cette relation dans le domaine de connaissances considéré. Le processus de validation dépend alors de la stratégie d'identification préalablement employée.

Le processus d'identification des relations peut ne nécessiter aucune étape de validation à proprement parler. C'est notamment le cas des approches manuelles qui ont recours à une expertise humaine, comme par exemple Terminae (Biebow & Szulman, 1999) qui confie à un expert la tâche (outillée) de sélectionner, dans une liste de termes candidats, ceux qui dénotent des concepts du domaine, puis de relier ces concepts pour former un réseau de termino-concepts. D'autres approches ont recours à des ressources sémantiques externes qui guident l'annotation des termes dénotant les concepts ou des termes dénotant les relations. Par exemple, l'exploitation conjointe du thesaurus UMLS [1] et de la Gene Ontology [2] en bioNLP permet de caractériser

1. https ://www.nlm.nih.gov/research/umls/
2. http ://geneontology.org/

les interactions entre gènes (Mcdonald *et al.*, 2004). Certaines approches adjoignent des patrons pour caractériser les relations ciblées (Embarek & Ferret, 2007).

L'étape de validation peut constituer une étape à part entière, suite au processus d'identification des relations. Il s'agit dans ce cas, soit de comparer les relations identifiées avec celles existant dans des ressources essentiellement lexicales (Wordnet et Eurowordnet sont largement utilisés à cet effet) (He & Da-You, 2004), soit de comparer les relations identifiées avec celles produites par un groupe d'annotateurs ayant obtenu un taux d'accord correct (> à 0.6 (Carletta, 1996)) suite à une campagne d'annotation (Mukherjee *et al.*, 2014), soit encore de comparer les relations avec un modèle de référence (certains sont disponibles par le biais des campagnes d'évaluation telles que BioNLP Shared Task 2011 (Kim *et al.*, 2011)). Dans ces contextes, les évaluations se font à l'aide de mesures comme *Taxonomy Overlap* (Cimiano *et al.*, 2005), qui compare le chevauchement entre deux taxonomies, ou celle qui mesure les correspondances en termes de *Coverage* (nombre de paires communes à la taxonomie générée et à la ressource correspondante), *Novelty* (nombre de relations correctes mais qui ne sont pas présentes dans la ressource) et *ExtraCoverage* (nombre de relations correctes mais sans correspondances dans la ressource sur le nombre total de relations dans la ressource)(Ponzetto & Strube, 2011).

Un autre type de validation concerne les approches endogènes, celles qui n'ont recours à aucune ressource externe. Une première approche, dite « par renforcement » ou « consolidation » nécessite de gros volumes de données, à l'échelle du web. La validation se base sur la redondance des informations issues de différentes sources, pour pouvoir confronter les résultats produits par différents extracteurs. Ces extracteurs peuvent être basés sur une approche symbolique (extracteur d'infobox, extracteur de tableau, etc.) dans Yago (Suchanek *et al.*, 2007), ou sur des techniques d'apprentissage (extracteur de table, extracteur de texte) dans Nell (Carlson *et al.*, 2010), ou plus précisément sur des classifieurs comme dans Google Knowledge Vault (Dong *et al.*, 2014). Ces approches de consolidation peuvent également mettre en œuvre des règles d'inférence pour détecter les similitudes et les inconsistances logiques (Suchanek *et al.*, 2009). Une autre approche, dite collaborative, permet d'attribuer un taux de confiance aux relations à travers des jeux sérieux (Lafourcade & Zampa, 2009) (voir par exemple « jeu de mots » à http ://www.jeuxdemots.org/).

Toutes ces approches souffrent cependant de limites : la vérification par rapport à des ressources externes se limite aux relations lexicales exprimées dans ces ressources, le recours aux experts ou aux annotateurs est très coûteux, il n'existe pas toujours de modèle de référence dans le domaine considéré, le corpus n'est pas toujours suffisamment volumineux. L'approche que nous proposons, bien qu'actuellement confinée aux SEP, pallie certains de ces inconvénients en n'exigeant aucune expertise humaine, en étant indépendante de la taille du corpus, et en offrant la possibilité de valider des relations qui ne sont pas totalement spécifiées dans les ressources utilisées. Mais avant de présenter l'approche proposée, nous rappelons dans la section suivante les caractéristiques et propriétés des SEP.

3 Structures énumératives parallèles

La structure énumérative (SE) est une structure textuelle ayant la propriété d'exprimer des connaissances hiérarchiques au travers de différents composants : une amorce, une liste d'items (au moins deux) constituant l'énumération, et éventuellement une clôture qui, lorsqu'elle existe, synthétise les différentes propositions exprimées à travers les items. Sur le plan sémantique, la

SE forme un tout. Sur le plan de la mise en forme, elle peut être exprimée selon différents modes, allant d'une forme linéaire sans mise en forme (Figure 1(a)) ou usant de caractères typographiques (Figure 1(b)), à une forme non linéaire usant de dispositifs typographiques et dispositionnels (Figure 1(c)).

(a) Le dromadaire a été répertorié dans 35 pays, tels que l'Inde, la Turquie, le Kenya, le Pakistan, la corne de l'Afrique et bien d'autres encore.

(b) Certaines plantes (palmiers, bambous. . .) produisent des tissus lignifiés.

(c) Une chaussure se compose principalement :
- du semelage, partie qui protège la plante des pieds
- de la tige, partie supérieure qui enveloppe le pied

FIGURE 1 – Différentes formes d'expression de la structure énumérative parallèle.

Plusieurs définitions de l'énumération existent, qui s'accordent sur l'égalité d'importance entre les différentes entités énumérées, par rapport à un critère de recensement (Dammame-Gilbert, 1989; Pascual, 1991). C'est toutefois la définition de Virbel (Virbel, 1989) qui semble le mieux prendre en compte à la fois les phénomènes architecturaux du texte et l'intention de l'auteur : "énumérer mobilise deux actes : un acte mental d'identification des éléments d'une réalité du monde dont on vise un recensement, et où on établit une relation d'égalité d'importance par rapport au motif de recensement ; et un acte textuel qui consiste à transposer textuellement la coénumérabilité des entités recensées, par la coénumérabilité des segments linguistiques qui les décrivent.".

La SE a également fait l'objet de nombreuses études au cours desquelles différentes typologies ont pu être proposées. Les SE linéaires ont été essentiellement analysées dans le cadre de l'analyse du discours. Elles ont donné lieu à des typologies comme celle de (Vergez-Couret *et al.*, 2008) où les SE à un temps ont été opposées aux SE à deux temps, ou encore comme celle de (Ho-Dac *et al.*, 2010) où les SE ont été classifiées selon leur niveau de granularité (SE dont les items sont des titres, SE en tant que listes formatées, SE multi-paragraphiques sans marque visuelle, SE intra-paragraphiques). Les SE usant de dispositifs typo-dispositionnels (dites verticales) ont quant à elles été notamment analysées dans le cadre de la génération de texte. (Hovy & Arens, 1998) distinguent les listes d'items (ensemble de composants de même niveau) des listes énumérées (pour lesquelles l'ordre des composants est pris en compte), alors que (Christophe, 2000) propose une typologie qui oppose les SE parallèles (paradigmatiques, homogènes visuellement et isolées) aux SE non parallèles. Cette dernière typologie est basée sur la composition du modèle rhétorique de la Rhetorical Structure Theory (RST) (Mann & Thompson, 1988) et du Modèle d'Architecture Textuelle (MAT) (Virbel, 1989).

Nous nous intéressons ici aux relations portées par les structures énumératives parallèles (SEP) au sens de Luc (2000), c'est à dire les SE pour lesquelles les items de l'énumération sont tous fonctionnellement équivalents (du point de vue syntaxique et rhétorique) et indépendants dans un contexte donné. Les exemples donnés ci-dessus correspondent à des SEP. D'un point de vue discursif, si la segmentation est faite en fonction des composants de la SE, c'est à dire que l'amorce et chacun des items correspondent à des unités de discours (UD), alors les UD relatives aux items sont successivement reliées par une relation rhétorique multi-nucléaire (ou

coordination), et l'UD relative à l'amorce est reliée à l'UD relative au premier item par une relation de type noyau-satellite (ou subordination). La figure 2 décrit le schéma discursif de la SEP selon la RST (Rhetorical Structure Theory).

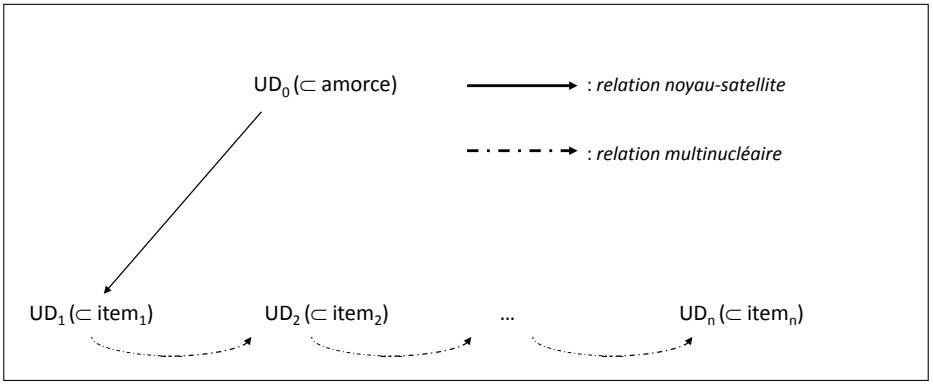


FIGURE 2 – Représentation discursive de la SE parallèle selon la RST (UD = Unité de Discours)

Selon les théories du discours (Asher, 1993), si " UD_j est subordonnée à UD_i, alors toute UD_k coordonnée à UD_j est subordonnée à UD_i". Ainsi, *N* relations de type noyau-satellite entre UD_0 et UD_i, pour *i=1,...,N* (si *N* est le nombre d'items dans la SE) peuvent être inférées. Ces *N* relations peuvent alors être spécialisées en *N* relations sémantiques $R(H, h_i)_{i=1,...,N}$ **de même nature**, où H correspond à un terme de UD_0, et h_i à un terme de UD_i.

Une SEP peut cependant porter plus de relations hiérarchiques, notamment lorsqu'un item est lui-même composé d'une énumération. Il est à noter qu'une SEP peut également porter des relations non hiérarchiques, dans ce cas généralement exprimées au sein de l'amorce ou d'un item. L'exemple de la Figure 3 illustre ces cas.

Les principaux actes cultuels sont :
- le sacrifice, la libation, l'offrande et l'éducation ;
- la prière (invocation, louange, demande, etc.) ;
- le chant et la musique ;
- la lecture de textes sacrés ;
- la prédication qui a un rôle important dans les religions abrahamiques ;
- les pèlerinages, processions.

FIGURE 3 – Exemple de SE où les items comportent des énumérations.

Cette SEP, outre le fait qu'elle intègre des SE sous diverses formes, montre que plusieurs relations hiérarchiques peuvent exister entre l'amorce et le premier item par exemple (*acte cultuel* et *sacrifice*, *acte cultuel* et *libation*, etc.), ainsi qu'une relation syntagmatique exprimée dans l'avant dernier item (*prédication* et *religions abrahamiques*).

Le travail proposé ici s'inscrit dans la continuité des travaux présentés dans (Fauconnier & Kamel, 2015), qui consistent à repérer les potentielles relations d'hyperonymie $R(H, h_i)_{i=1,...,N}$

au sein d'une SEP, à l'aide d'un système d'apprentissage supervisé. Dans ce contexte, la structure discursive de la SEP a été exploitée pour faire valoir la présence d'une relation hiérarchique entre un terme présent dans l'amorce et des termes présents dans les items (un terme par item). L'approche de validation que nous proposons ici exploite également la structure discursive des SEP pour faire valoir cette fois la proximité sémantique, notamment en termes de cohésion lexicale, entre les h_i (*i=1,...,N*) liés par une même relation à une même entité H.

4 Approche proposée

Les erreurs d'identification de relations par les systèmes d'extraction automatique de relations à partir de texte sont généralement dues soit à une mauvaise interprétation de la nature de la relation (difficulté pour les systèmes à résoudre certains phénomènes linguistiques comme les anaphores, les ellipses, etc.), soit à la mauvaise identification des arguments (difficultés à extraire les bons termes notamment). Une autre cause peut être la mauvaise formalisation des connaissances de l'auteur du texte. Il s'agit alors de ne valider que les relations qui portent sens.

4.1 Principe général

Le principe de validation que nous mettons en œuvre exploite les propriétés discursives de la SEP pour valider conjointement (et non indépendamment les unes des autres) les relations $R(H, h_i)$ (i=1,..,N) issues d'une même SEP, et où R correspond à la relation d'hyperonymie, H à l'hyperonyme, et h_i à l'hyponyme. Nous utilisons pour cela deux ressources sémantiques : un réseau lexico-sémantique qui fournit en général une bonne précision mais dont le taux de couverture est variable, et une ressource distributionnelle qui ne spécifie pas les relations entre termes mais dont la couverture est très large. Le principe de validation, appliqué à une SEP, se déroule en deux étapes :

1. valider toutes les relations $R(H, h_i)$ qui sont exprimées dans le réseau sémantique, avec un coefficient de 1.
2. valider toutes les relations $R(H, h_k)$ qui ne sont pas exprimées dans le réseau sémantique en évaluant le coefficient de proximité distributionnelle qu'entretient h_k avec tous les h_i appartenant aux relations validées à l'étape 1.

Nous décrivons ci-dessous le principe algorithmique de façon plus détaillée.

4.2 Algorithme

Nous donnons les définitions suivantes :
— RS le réseau sémantique, et RD la ressource distributionnelle ;
— R_i une relation candidate d'hyperonymie issue d'une SEP et liant l'hyperonyme H détecté dans l'amorce, à l'hyponyme h_i identifié dans le i^{me} item (*i=1,..., N* si *N* est le nombre d'items présents dans la SEP) ;
— ***Lexicalisation(C)*** l'ensemble des termes associés au concept C au sein de RS
— $Synset(H) = \{C \in RS / H \in Lexicalisation(C)\}$;
— $Hyperonymes_{RS}(h_i)$ l'ensemble des hyperonymes directs de h_i dans RS

— $SuperHyperonymes_{RS}(h_i)$ l'ensemble des hyperonymes de h_i existant dans RS et liés à h_i par un chemin de longueur inférieur ou égal à un seuil S fixé de façon empirique selon le réseau sémantique utilisé :
$SuperHyperonymes_{RS}(h_i) = \bigcup_{k=1}^{S} SuperHyperonymes_{RS}^{k}(h_i)$ où
$SuperHyperonymes_{RS}^{k}(h_i)$ est l'ensemble des hyperonymes de h_i de rang k (k étant la longueur maximale du chemin reliant h_i à un de ses hyperonymes dans RS), et où
$SuperHyperonymes_{RS}^{k}(h_i)$ est défini récursivement par :
$SuperHyperonymes_{RS}^{1}(h_i) = \{hyperonymes_{RS}(h_i)\}$;
$SuperHyperonymes_{RS}^{k}(h_i) = SuperHyperonymes_{RS}^{k-1}(h_i) \bigcup \bigcup_{h \in SuperHyperonymes_{RS}^{k-1}(h_i)} hyperonymes_{RS}(h)$.

Nous donnons la description complète de l'algorithme ci-dessous (Algorithm 1).

Algorithm 1 Principe de validation des relations issues d'une SEP

V est l'ensemble des relations validées
$\overline{V}$ est l'ensemble des relations non validées R_i
% au départ, toutes les relations ont le statut de "non validée" %
$V \leftarrow \emptyset$
$\overline{V} \leftarrow \bigcup_{i=0}^{N} R_i$
Pour chaque relation $R_i(H, h_i) \in \overline{V}$ **Faire**
  Si $SuperHyperonymes_{RS}(h_i) \cap Synset(H) \neq \emptyset$ **alors**
    $Valider(R_i(H, h_i)) \leftarrow 1$
    $V = V \bigcup \{R_i\}$
    $\overline{V} = \overline{V} - \{R_i\}$
  Fin Si
Fin Pour
% si au moins une des relations a obtenu le statut de "validée" (ce qui confirme l'hyperonymie), et si au moins une relation possède toujours le statut de "non validée" (ce qui amène à exploiter la ressource distributionnelle)%
Si $V \neq \emptyset$ *et* $\overline{V} \neq \emptyset$ **alors**
  Pour chaque relation $R_k(H, h_k) \in \overline{V}$ **Faire**
    $valider(R_k) = \frac{\sum_{R_i \in V} p(h_i, h_k)}{|V|}$ *% où p(h_i,h_k) correspond à la proximité sémantique fournie par RD %*
    % calcul de la somme des mesures de proximité entre l'hyponyme de la relation non validée h_k et les hyponymes h_i des relations validées %
  Fin Pour
Fin Si

5 Application et évaluation

5.1 Jeu de données

Le jeu de données que nous avons utilisé pour notre expérimentation est constitué de 262 relations candidates d'hyperonymie fournies par le système d'extraction de relations décrit dans (Fauconnier & Kamel, 2015) (comme dit précédemment, l'approche de validation que nous proposons s'inscrit dans la suite de ces travaux). Ce système extrait par apprentissage supervisé des relations d'hyperonymie à partir de SEP, ces SEP ayant été elles-mêmes extraites automatiquement de pages Wikipedia. En effet, les articles Wikipédia sont rédigés selon le guide "the Manual of Style" (`http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style`) qui préconise l'utilisation de SEP et recommande pour ces structures d'utiliser la même forme grammaticale pour tous les items. Une validation de ces structures énumératives en tant que parallèles avait été menée manuellement.

Les 262 relations composant notre jeu de données proviennent donc de 67 de ces SEP. Ces relations ont fait l'objet d'une validation manuelle effectuée par deux annotateurs en double aveugle, à la suite de laquelle 27 conflits inter-annotateurs ont été identifiés et résolus (i.e., les deux annotateurs se sont *a posteriori* mis d'accord). Les désaccords étaient principalement liés au fait que certaines relations pouvaient être caractérisées d'hyperonymie ou d'holonymie. Ainsi 206 relations ont été évalués comme étant correctes et 56 comme incorrectes. Ces relations constituent alors un ensemble de relations de référence.

5.2 Ressources utilisées

Nous avons utilisé le réseau lexico-sémantique *BabelNet* (Navigli & Ponzetto, 2012) et la ressource distributionnelle *Voisins de Wikipédia*[3] (Adam *et al.*, 2013). Le réseau *BabelNet* a été construit automatiquement à partir de l'intégration de plusieurs ressources (WordNet, Open Multilingual WordNet, Wikipedia, GeoNames, WoNef, etc.). Il est composé d'environ 14 millions d'entrées, incluant concepts et entités nommées, chaque entrée définissant un *BabelSynset*. Chaque *BabelSynset* correspond à un sens donné (*BabelSense*) et regroupe tous les synonymes dans 271 langues différentes, dont la langue française. La ressource *Voisins de Wikipédia* a éte construite à partir d'un corpus de 262 millions de mots, selon les principes décrits par (Bourigault, 2002) à partir d'un modèle structuré (Baroni & Lenci, 2010).

Ces ressources ont été choisies d'une part car elles expriment des connaissances en langue française, et d'autre part car elles ont pour origine le même corpus que celui dont est issu le jeu de données.

5.3 Conditions de l'expérimentation

Nous avons fixé de façon empirique la longueur maximale du chemin reliant un hyponyme à l'un de ses hyperonymes à k=3. Par ailleurs, les mesures de confiance accordées aux relations validées par le système ont été calculées de la façon suivante :

— nous avons considéré que toute relation entre un hyperonyme et un de ses hyponymes exprimée dans *BabelNet* a une mesure de confiance égale à 1.

3. http ://redac.univ-tlse2.fr/applications/vdw.html

— les mesures de confiance entre deux hyponymes sont celles données par la ressource distributionnelle *Voisins de Wikipedia*. Ces mesures correspondant à des scores de Lin (Lin, 1998) compris entre 0,1 et 0,29 pour 97% des entrées (Adam *et al.*, 2013).

5.4 Résultats

Nous avons expérimenté et évalué notre approche sur deux ensembles de relations candidates :

— E l'ensemble des relations du *gold standard* (206 relations)

— E_{BN} l'ensemble des relations dont l'hyperonyme possède une entrée dans *BabelNet* (116 relations)

L'évaluation a été menée en termes de précision, rappel et exactitude. Ces résultats sont présentés dans la Table 1.

	Précision	Rappel	F-Measure	Exactitude
E	.97	.37	.54	.50
E_{BN}	.97	.66	.78	.67

TABLE 1 – Résultats de la validation des relations candidates.

Nous avons obtenu le même taux de précision pour les deux ensembles E et E_{BN} : 76 relations parmi les 78 relations automatiquement validées sont correctes. Comme attendu, nous avons obtenu de moins bons résultats en termes de rappel. Pour l'ensemble E, 76 relations ont été validées sur les 206 correctes. Les résultats pour l'ensemble E_{BN} sont meilleurs, 76 relations ont été validées sur 116 correctes. En termes d'exactitude, pour l'ensemble E, parmi les 262 relations (impliquant les relations annotées comme correctes et incorrectes par les annotateurs), 130 ont été correctement validées par le système. Pour l'ensemble E_{BN}, 88 relations ont été correctement confirmées, sur un total de 131.

Il est à noter que 12 relations parmi les 76 validées par le système, ont été validées par la ressource distributionnelle, et sont correctes. Dans ce contexte, la ressource distributionnelle a validé nos relations avec une précision de 1.0, et a permis d'améliorer les performances de notre système à hauteur de 15%.

5.5 Discussion

Bien que la précision soit très haute, nous avons pu identifier dans quel cas notre système validait une relation fausse : cela est dû au fait d'utiliser des BabelSynsets qui regroupent des termes de sens proche. Par exemple, lors de l'évaluation de la relation candidate $R(pays, Corne\ de\ l'Afrique)$, le BabelSynset $bn\ :\ 00028934n$ composé par les BabelSenses {terre, sol, terre ferme, contrée, pays} appartient à l'intersection des ensembles $SuperHyperonymes^{3}_{BN}(Corne\ de\ l'Afrique)$ et $Synset(pays)$.

Pour ce qui concerne le rappel, nous avons identifié deux causes au silence. La première cause est une conséquence de l'absence de l'hyperonyme comme entrée dans *BabelNet* (e.g., 'feuillus colonisateur' ou 'poisson sauvage'). Dans ce cas, aucune relation de la SEP porteuse de cet hyperonyme n'est validée. Cela concerne 62 relations pour l'ensemble E. La deuxième

cause provient du fait que la valeur fixée à 3 pour la profondeur du chemin de recherche des hyperonymes de l'hyponyme h_i n'est quelquefois pas suffisante. Augmenter alors la valeur de k permettrait d'améliorer le rappel, mais nécessiterait de définir des heuristiques pour conserver une complexité algorithmique acceptable.

La ressource distributionnelle permet d'identifier des relations non exprimées dans le réseau lexico-syntaxique. Par exemple, la relation *R(anomalie chromosomique,insertion)* absente de *BabelNet* a pu être validée, par le fait que *R(anomalie chromosomique,délétion)* est présente dans *BabelNet*, et que les termes *insertion* et *délétion* sont sémantiquement proches dans la ressource distributionnelle. Bien que les entrées de cette ressource ne correspondent très majoritairement qu'à des mots simples, et non à des termes, alors que 40% des hyponymes de relations que nous avions à valider correspondent à des termes composés de plus d'un mot, les performances par rapport à l'exploitation du réseau lexico-syntaxique seul ont pu être améliorées de 15%. Une ressource distributionnelle intégrant les termes permet d'envisager d'améliorer encore ces résultats.

6 Conclusion et Perspectives

Ce travail, qui s'inscrit dans la continuité de travaux entrepris sur l'identification de relations d'hyperonymie portées par les structures énumératives parallèles, a pour objet de proposer une méthode de validation automatique de ces relations. L'originalité de l'approche proposée tient au fait qu'elle exploite la structure discursive de la structure énumérative parallèle, et qu'elle met l'accent sur la complémentarité entre deux ressources de nature différente, un réseau lexico-sémantique et une ressource distributionnelle. L'approche a été évaluée sur les SEP, avec une exactitude comprise entre 0.5 et 0.67 selon les conditions expérimentales, et avec une amélioration de 15% des performances due à l'exploitation conjointe de la ressource distributionnelle. Cette approche vaut pour tout objet textuel ayant même schéma discursif que la SEP, comme les titres et sous-titres, les champs de formulaire, etc.

Une des premières suites que nous prévoyons à ce travail est de voir comment l'exploitation des ressources externes peut conduire à améliorer notre système. Plusieurs pistes sont envisagées : ressources distributionnelles intégrant les termes (idéalement du domaine), analyser le compromis entre étendue des recherches au sein du réseau sémantique vs. complexité algorithmique, d'autres réseaux sémantiques voire utiliser conjointement plusieurs réseaux lexico-syntaxiques et plusieurs ressources distributionnelles.

Une autre suite envisagée concerne l'adaptation et l'évaluation de notre approche pour des relations lexicales autres que l'hyperonymie, comme la méronymie, la synonymie et l'antonymie.

Remerciements

Cassia Trojahn est partiellement financée par le projet FUI SparkinData.

Références

ADAM C., FABRE C. & MULLER P. (2013). Évaluer et améliorer une ressource distributionnelle : protocole d'annotation de liens sémantiques en contexte. *TAL*, **54**(1), 71–97.

ASHER N. (1993). *Reference to Abstract Objects in Discourse : A Philosophical Semantics for Natural Language Metaphysics*, volume 50 of *SLAP*. http ://www.wkap.nl/ : Kluwer.

BARONI M. & LENCI A. (2010). Distributional memory : A general framework for corpus-based semantics. *Comput. Linguist.*, **36**(4), 673–721.

BIEBOW B. & SZULMAN S. (1999). Terminae : A linguistic-based tool for the building of a domain ontology. In *Proceedings of the 11th European Workshop on Knowledge Acquisition, Modeling and Management*, EKAW '99, p. 49–66, London, UK, UK : Springer-Verlag.

BOURIGAULT D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de la 9e conférence sur le Traitement Automatique de la Langue Naturelle*, TALN'2002, p. 75–84.

BUITELAAR P., CIMIANO P. & MAGNINI B. (2005). *Ontology Learning from Text : An Overview*, In P. BUITELAAR, P. CIMIANO & B. MAGNINI, Eds., *Ontology Learning from Text : Methods, Evaluation and Applications*, volume 123. IOS Press.

CARLETTA J. (1996). Assessing agreement on classification tasks : The kappa statistic. *Comput. Linguist.*, **22**(2), 249–254.

CARLSON A., BETTERIDGE J., KISIEL B., SETTLES B., JR. E. R. H. & MITCHELL T. M. (2010). Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*.

CHRISTOPHE L. (2000). *Représentation et composition des structures visuelles et rhétoriques du textes. Approche pour la génération de textes formatés*. PhD thesis.

CIMIANO P., HOTHO A. & STAAB S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. *J. Artif. Int. Res.*, **24**(1), 305–339.

DAMMAME-GILBERT B. (1989). *La série énumérative : étude linguistique et stylistique s' appuyant sur dix romans français publiés entre 1945 et 1975*, volume 19. Librairie Droz.

DONG X., GABRILOVICH E., HEITZ G., HORN W., LAO N., MURPHY K., STROHMANN T., SUN S. & ZHANG W. (2014). Knowledge vault : A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, p. 601–610, New York, NY, USA : ACM.

EMBAREK M. & FERRET O. (2007). Une expérience d'extraction de relations sémantiques à partir de textes dans le domaine médical. In *Proceedings of 14ème Conférence sur le Traitement automatique des langues naturelles (TALN 2007)*, p. 37–46.

FAUCONNIER J.-P. & KAMEL M. (2015). Discovering Hypernymy Relations using Text Layout (regular paper). In *Joint Conference on Lexical and Computational Semantics (SEM), Denver, Colorado, 04/06/15-05/06/15*, p. 249–258, http ://www.aclweb.org : Association for Computational Linguistics (ACL).

HE H. & DA-YOU L. (2004). Learning owl ontologies from free texts. In *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on (Volume :2)*, p. 1233–1237 : IEEE Conference Publications.

HO-DAC L.-M., PÉRY-WOODLEY M.-P. & TANGUY L. (2010). Anatomie des Structures Énumératives. In *Traitement Automatique des Langues Naturelles*, p. (publication numérique), Montréal, Canada.

HOVY E. H. & ARENS Y. (1998). Readings in intelligent user interfaces. chapter Automatic Generation of Formatted Text, p. 256–262. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

KIM J.-D., PYYSALO S., OHTA T., BOSSY R., NGUYEN N. & TSUJII J. (2011). Overview of bionlp shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, BioNLP Shared Task '11, p. 1–6, Stroudsburg, PA, USA : Association for Computational Linguistics.

LAFOURCADE M. & ZAMPA V. (2009). Pticlic : a game for vocabulary assessment combining jeux-demots and lsa. In *In proc of CICLing (Conference on Intelligent text processing and Comptational*

Linguistics). Mexico : Marsh 1-7.

LIN D. (1998). An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, p. 296–304, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

MANN W. C. & THOMPSON S. A. (1988). Rhetorical structure theory : Toward a functional theory of text organization. *Text*, **8**(3), 243–281.

MCDONALD D. M., CHEN H., SU H. & MARSHALL B. B. (2004). Extracting gene pathway relations using a hybrid grammar : the arizona relation parser. *Bioinformatics*, p. 3378.

MUKHERJEE S., AJMERA J. & JOSHI S. (2014). Domain cartridge : Unsupervised framework for shallow domain ontology construction from corpus. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, p. 929–938.

NAVIGLI R. & PONZETTO S. P. (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, **193**, 217–250.

PASCUAL E. (1991). *Représentation de lárchitecture textuelle et génération de texte*. PhD thesis.

PONZETTO S. & STRUBE M. (2011). Taxonomy induction based on a collaboratively built knowledge repository. volume 9 of *175*, p. 1737–1756.

SUCHANEK F. M., KASNECI G. & WEIKUM G. (2007). Yago : A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, p. 697–706, New York, NY, USA : ACM.

SUCHANEK F. M., SOZIO M. & WEIKUM G. (2009). Sofie : a self-organizing framework for information extraction. In J. QUEMADA, G. LEÓN, Y. S. MAAREK & W. NEJDL, Eds., *WWW*, p. 631–640 : ACM.

VERGEZ-COURET M., PRÉVOT L. & BRAS M. (2008). Interleaved discourse, the case of two-step enumerative structures. p. 85–94 : Proceedings of Contraints In Discourse III, Postdam.

VIRBEL J. (1989). Structured documents. chapter The Contribution of Linguistic Knowledge to the Interpretation of Text Structures, p. 161–180. New York, NY, USA : Cambridge University Press.

Extraire semi-automatiquement des connaissances dans la littérature biomédicale

Jessica Pinaire[1,3], Jérôme Azé[1], Sandra Bringay[1,4], Paul Landais[2,3]

[1] LIRMM, UMR 5506, Université de Montpellier, France
prenom.nom@lirmm.fr

[2] ÉQUIPE D'ACCUEIL 24-15, Institut Universitaire de Recherche Clinique, Université de Montpellier, Montpellier, France
paul.landais@inserm.fr

[3] CHU, Département d'information médicale, Nîmes, France
paul.landais@chu-nimes.fr

[4] AMIS, Université Paul Valéry, Montpellier, France
sandra.bringay@univ-montp3.fr

Résumé :
La littérature biomédicale résume les connaissances scientifiques actuelles. La quantité de données disponibles est trop importante pour pouvoir être analysée manuellement. Il devient crucial de construire des outils d'analyse automatisés pour supporter les activités de recherche bibliographique. Dans ce contexte, nous proposons une méthode semi-automatique, supportée par des visualisations, pour explorer les bases bibliographiques, que nous appliquons à la thématique des trajectoires de patients.

Nous avons utilisé et évalué plusieurs modèles statistiques et leurs représentations visuelles pour caractériser le contenu des articles et donner au chercheur une vision d'ensemble du contenu de la base.

L'approche a permis d'identifier 81 articles dans la littérature parmi 11 490 articles, qui ont été étudiés lors d'une revue systématique manuelle.

Nous proposons une approche semi-automatique efficace pour l'exploration de la littérature biomédicale.

Mots-clés : Fouille de textes, Visualisation de connaissances, Recherche d'informations.

1 Introduction et motivations

Notre travail se situe dans le contexte de l'aide à la réalisation d'études bibliographiques. Notre intérêt pour le domaine d'application biomédical est motivé par le fait que ce domaine a connu la plus forte croissance de tous les domaines scientifiques en terme de volume de publications. En février 2016, le moteur de recherche PubMed indexe plus de 25 millions de citations. Si plusieurs communautés (*e.g.* recherche d'informations, fouille de textes) se sont penchées sur le défi de l'extraction automatisée d'informations dans la littérature (Fleuren & Alkema, 2015; Huang & Lu, 2016; Vazquez *et al.*, 2011; Song, 2014; Geifman *et al.*, 2015), à ce jour, il n'existe pas, à notre connaissance, d'outils permettant aux chercheurs d'explorer facilement ces grands volumes de données.

Dans cet article, nous proposons de combiner plusieurs outils de fouille de textes pour aider le chercheur à visualiser le contenu de très nombreux articles en mettant en évidence les thèmes abordés, à l'aide de mots saillants et de réseaux de mots. L'objectif de ces visualisations est d'aider l'expert à formuler une question de recherche précise qui lui permettra dans un deuxième temps de filtrer le corpus et d'aboutir à un nombre de documents manuellement exploitables pour réaliser une revue systématique classique. Cette méthode a été appliquée avec succès pour une étude sur le thème des trajectoires de patients.

La section 2 motive ces travaux et décrit un rapide état de l'art. La section 3 décrit la stratégie employée et les outils mis en œuvre. La section 4 détaille la partie applicative sur le corpus de textes des trajectoires. Finalement, nous concluons et donnons quelques perspectives.

2 Motivations et bref état de l'art

Un travail de recherche doit nécessairement débuter par une analyse des bases bibliographiques spécialisées, afin de positionner ces travaux par rapport à l'état des connaissances scientifiques reflété via les publications.

Dans ce contexte, une revue systématique correspond à une recherche bibliographique basée sur une question clairement formulée. Dans un premier temps, une telle revue utilise des méthodes systématisées, répétitives et explicites pour identifier, sélectionner et évaluer de façon critique des articles de recherche répondant à cette question. Dans un deuxième temps, la revue permet de recueillir et d'analyser les données extraites des publications filtrées. Une méta-analyse peut alors être utilisée en complément d'une revue pour analyser et résumer les résultats des études. Cette méta-analyse se base sur des techniques statistiques. On trouve plusieurs méthodes de revues systématiques comme la méthode PRISMA [1] (Moher *et al.*, 2009) ou Cochrane [2] (Higgins *et al.*, 2008). Si ces méthodes s'avèrent très efficaces quand la question de recherche est clairement formulée, elles ne donnent que peu de directives quand le besoin informationnel du chercheur est plus vague.

C'est justement dans ce contexte que nous nous positionnons. Notre objectif est de proposer une méthode semi-automatique, facilitant le processus de recherche bibliographique en permettant une exploration sans *a priori* des articles. Notre objectif est d'aider le chercheur à identifier les thèmes d'intérêts de sa communauté, de se focaliser sur certains et de formuler une série de questions de recherche plus précises, qui pourront ensuite être utilisées en entrée d'une revue automatique.

Nous prenons ici comme cas d'étude la thématique des "trajectoires de patients". Ce champ d'étude a donné lieu à un nombre croissant de publications scientifiques dans de nombreuses revues médicales au cours des dix dernières années.

Nous utilisons le moteur de recherche PubMed développé par le NCBI [3] et considéré comme l'une des références dans les domaines de spécialisation de la biologie et de la médecine. Il donne accès à la base de données bibliographique MEDLINE qui contenait en février 2016 plus de 25 millions de citations publiées depuis 1950 dans environ 5 000 revues biomédicales [4]. Des travaux ont démontré que MEDLINE possède la couverture la plus complète des références bibliographiques biomédicales (Kelly & St Pierre-Hansen, 2008). PubMed est largement utilisé par la communauté scientifique biomédicale. En 2015, il y a eu, en moyenne, près de 8 millions de requêtes par jour [5].

1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses
`http://www.prisma-statement.org/`
2. `http://community.cochrane.org/cochrane-reviews`
3. National Center for Biotechnology Information
4. `http://www.ncbi.nlm.nih.gov/pubmed`
5. `https://www.nlm.nih.gov/services/pubmed_searches.html`

Pour notre cas d'étude, si l'on interroge pubmed avec les mots clés "health", "patient(s)", "trajectories", "trajectory", "path", "pathway(s)", sur une période de presque 15 ans, nous obtenons un corpus de 11 490 documents, qu'il est impossible d'exploiter manuellement. Cet exemple montre que développer des méthodes pour un accès intelligent aux données contenues dans l'ensemble des publications scientifiques biomédicales reste un véritable défi.

Dans le domaine de la recherche bibliographique, les méthodes de fouilles de textes ont été utilisées pour faciliter le travail du chercheur en ciblant mieux la recherche documentaire et en réduisant le temps de travail (Cohen *et al.*, 2006). Dans le cadre de review systématiques, par exemple, il existe quatre tâches (Thomas *et al.*, 2011) pour lesquelles les techniques de fouille de textes sont généralement employées : 1) la reconnaissance automatique de termes dans les textes (Frantzi *et al.*, 2000) ; 2) la classification supervisée de documents dans des thèmes spécifiques (Joachims, 1998; Mo *et al.*, 2015; Sebastiani, 2002; Frunza *et al.*, 2011) ; 3) la classification non supervisée de documents qui regroupe les documents dans des thèmes. Chaque groupe correspond au thème partagé par l'ensemble des documents du groupe et par aucun autre document de la collection (Blei *et al.*, 2003; Reinert, 1983; Bada, 2014) ; 4) le résumé fait en sélectionnant des phrases à partir de chaque document basé sur l'importance de ses termes qui sont combinés avec des techniques de classement (Bollegala *et al.*, 2010).

D'autres auteurs utilisent la fouille de textes pour d'autres finalités. Par exemple, dans (Lin *et al.*, 2008), les auteurs créent des bases de données de correspondances reliant les auteurs avec les abréviations de leurs noms et réalisent une analyse des co-auteurs. Dans (Leitner & Valencia, 2008), ils annotent les abstracts de deux façons, d'abord le gène ou la protéine étudiée, puis les interactions de protéines et/ou les fonctions du gène. *In fine*, ils catégorisent les documents suivant ces annotations.

Si ces méthodes sont intéressantes, elles ne permettent pas d'explorer de manière globale et sans *a priori*, c'est-à-dire sans question de recherche précisément définie, les grands corpus d'articles. Notre méthode permet non seulement une telle exploration visant à filtrer les documents pertinents pour une revue plus systématique mais également d'accompagner cette exploration par des visualisations qui facilitent l'interprétation du chercheur. Dans la section suivante 3, nous décrivons les différentes étapes de cette méthode.

3 Analyse bibliographique semi-automatisée par fouille de textes

Notre méthode semi-automatique, basée sur des techniques de fouille de textes, aide : 1) à interpréter de gros volumes de données générés par la littérature biomédicale ; 2) à explorer le corpus par raffinements successifs jusqu'à la formulation d'une question de recherche. Nous détaillons dans cette section les trois étapes de la méthode (voir figure 1).

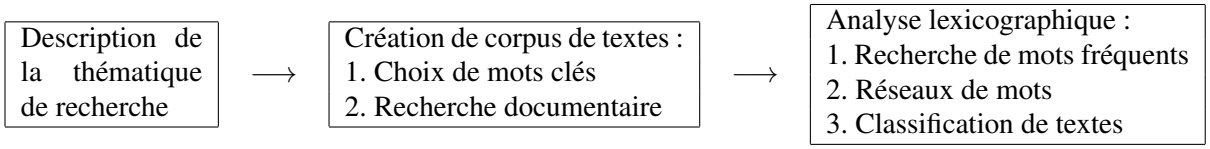


FIGURE 1 – Stratégie d'analyse

3.1 Étape 1 : Définition de la thématique de recherche

Afin de débuter l'exploration du corpus, nous demandons à l'expert de définir sa thématique de recherche en langage naturel en explicitant son besoin informationnel. Cela nous permet de définir des questions sans *a priori* qui intègrent des aspects thématiques *e.g. quelles communautés travaillent sur ma thématique ? Quelles maladies sont concernées par cette thématique ? Quels sont les sujets abordés par les articles ?*

Dans un deuxième temps, l'expert doit choisir une liste de mots clés pour représenter cette thématique qui sera utilisée à l'étape 2 pour générer le corpus.

3.2 Étape 2 : Création du corpus de textes

À partir du premier ensemble de mots clés, nous utilisons l'API de Pubmed pour rechercher les documents. Un corpus de textes, constitué de l'union du titre et de l'abstract est ensuite créé. Afin de ne retenir que les autres termes, les mots clés utilisés pour la sélection sont supprimés du corpus précédemment créé. Ensuite, une première exploration du corpus par Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003), à l'aide du package LDAviz (Sievert & Shirley, 2014) dans R, permet de repérer des abréviations (*e.g.* ppci, rr, hr) peu répandues dans le langage courant, que nous écartons du corpus.

Enfin, les prétraitements suivants sont appliqués : a) lemmatisation du texte : les verbes sont ramenés à l'infinitif, les noms au singulier et les adjectifs au masculin singulier ; b) enrichissement du dictionnaire : le logiciel détecte des termes non reconnus. Pour ne pas perdre trop d'informations, ces termes non reconnus ont été récupérés, lemmatisés par TreeTagger [6] et réinjectés dans le dictionnaire après vérification manuelle et ajout de termes médicaux spécifiques ou sigles bien connus comme AMI (Acute Myocardial Infarction). Dans la suite, les analyses sont réalisées avec les formes pleines (noms, adjectifs, adverbes et verbes).

3.3 Étape 3 : Analyse lexicographique

Nous analysons le corpus précédent avec le logiciel IRaMuteQ [7]. Il s'agit d'une Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires (Ratinaud & Déjean, 2009). Il permet de faire des analyses statistiques sur des corpus de textes (Ratinaud & Marchand, 2012).

Nuage de mots : C'est une représentation synthétique de la distribution des termes. Les mots les plus fréquents sont au centre et la taille de police varie proportionnellement au nombre d'occurrences.

Analyse de similarités : C'est une technique, reposant sur la théorie des graphes, classiquement utilisée pour décrire des représentations sociales, sur la base de questionnaires d'enquête (Flament, 1981). L'objectif de l'analyse de similarités (ADS) est d'étudier la proximité et les relations entre les éléments d'un ensemble, sous forme d'"arbres maximum". Pour chaque corpus, nous avons choisi la représentation en arbre *graphopt* décrite dans (Csardi, 2015) et l'algorithme de (Brandes, 2001) pour décrire les communautés au sens du plus court chemin, ce qui met en évidence les mots les plus souvent associés dans une même phrase ou un texte.

6. `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

7. `http://www.iramuteq.org/`

Classification de textes : La classification de Reinert (Reinert, 1983) est une classification hiérarchique descendante (CDH) s'effectuant en plusieurs étapes. Elle propose une approche globale du corpus. Après partitionnement de celui-ci, elle identifie des classes statistiquement indépendantes de mots. Ces classes sont interprétables grâce à leurs profils, qui sont caractérisés par des mots spécifiques corrélés entre eux. La CDH résume cela par un dendrogramme.

4 Application au corpus Trajectoire

Le CHU de Nîmes s'intéresse aux trajectoires de patients. Dans cette section, nous allons appliquer notre méthode en collaboration avec un expert de cette thématique. L'expert est un professeur de médecine spécialisé dans l'analyse des bases médico-économiques.

4.1 Étape 1 : Définition de la thématique de recherche

La table 1 décrit les questions sans *a priori* formulées par le chercheur.

Questions sans *a priori*
Q1 : Existe-t-il des études sur les trajectoires de patients ?
Q2 : Quels sont les thèmes abordés dans ces études ? (la prise en charge, le traitement, les coûts,...)
Q3 : Pour quelles pathologies sont étudiées les trajectoires ?

TABLE 1 – Questions sans *a priori*

À partir de ces questions, l'expert choisi des mots clés décrivant la thématique qu'il souhaite explorer. Nous retenons les termes suivants pour décrire la notion de trajectoire : "trajectoire", "parcours" et "chemin".

4.2 Étape 2 : Création du corpus de texte

Dans PubMed, nous avons recherché les documents qui traitent des trajectoires ou parcours de soins de patients. Nous avons effectué une recherche selon le thème et contraintes résumés dans le tableau 2. Ainsi, la requête : **C1 + T1 + C2 + C3**, sélectionne les articles du domaine médical qui traitent des trajectoires, écrits entre le 1^{er} janvier 2000 et le 31 octobre 2015, en anglais. Il a résulté de la recherche documentaire un total de 11 490 articles.

Thème et contraintes	**Mots clés**
C1 : contexte medical	"health", "patient(s)"
T1 : Trajectoire	"trajectories", "trajectory", "path", "pathway(s)"
C2 : dates	January 1st 2000 to October 31th 2015
C3 : langue	English

TABLE 2 – Mots clés utilisés dans la recherche documentaire

4.3 Étape 3 : Analyse Lexicographique

Nuage de mots : Les termes les plus saillants de la figure 2 sont "care", "study", "cancer", "cell", "treatment", "disease". Les termes "study" et "care" nous apportent des éléments sur le contenu du corpus : ces articles traitent de la notion de soins. Ensuite, avec les termes, "treatment", "disease", il est avant tout question de parcours de soins du patient. Nous pouvons alors répondre positivement à la question **Q1** : il existe bien des études sur les trajectoires de patients.

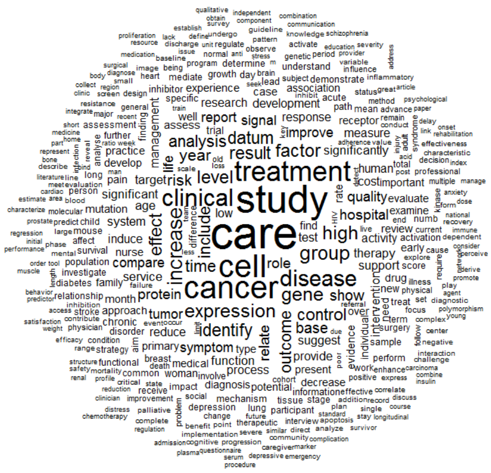


FIGURE 2 – Nuage de mots du corpus Trajectoire

Cette première analyse fait ressortir les mots importants caractéristiques du corpus. Toutefois, au delà de l'intérêt visuel, le nuage n'apporte pas beaucoup d'informations. Avec l'ADS, nous allons explorer en détail ce corpus, en rapprochant ces termes d'autres termes pour en déduire leur signification et définir un contexte.

Analyse de similarités : Cette présentation en arborescence met en évidence des réseaux de termes fortement co-occurents et permet ainsi de cerner les thématiques les plus souvent abordées au travers des articles. La figure 3 montre que l'étude est fortement liée au soin, à la maladie et plus spécifiquement au cancer. Dans la représentation des communautés, on voit apparaître celle des maladies étudiées, causant des dysfonctionnements sévères, chroniques, du cœur, des reins ou encore des poumons, que l'on soit jeune, âgé, homme ou femme. Ces deux dernières remarques répondent à la question **Q3**.
Notons que la notion de cancer est représentée dans une feuille différente de celle des maladies. Cette observation permet à l'expert de se poser la question suivante : *pour quelles raisons cette pathologie est-elle particularisée dans le contexte des trajectoires ?* Cette feuille contient le mot "treatment", qui est une notion intrinsèque à celle de trajectoire car cohérent avec le travail réalisé par les chercheurs de cette discipline ces dernières années pour organiser les soins de cette pathologie lourde.

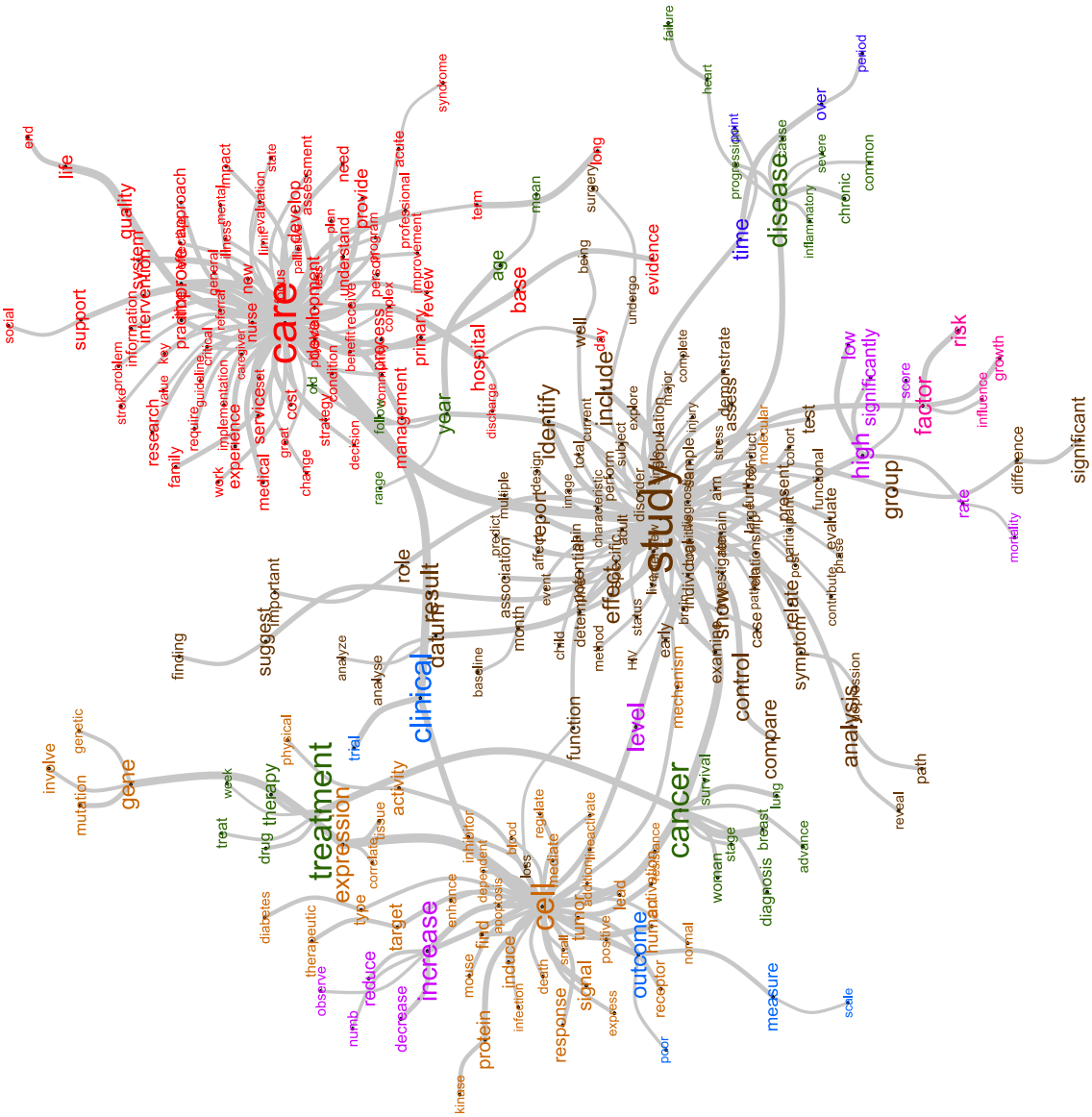


FIGURE 3 – ADS et communautés du corpus Trajectoire

Dans une autre branche la notion de cancer est liée à celle de génétique. L'utilisation du mot clé "pathway" fait ressortir tous les articles évoquant les voies de signalisation cellulaire et les gènes. Dans notre contexte applicatif, cette visualisation nous permet d'éliminer ce thème car il ne s'agit plus de la trajectoire du patient au sens série d'événements médicaux que nous souhaitons étudier mais d'interactions au niveau des cellules. De plus, ceci montre que plusieurs définitions de la notion de trajectoire existent et que le sens doit être affiné via un item à ajouter dans la grille de lecture de la revue systématique.

L'ADS a permis d'explorer en partie le contenu des articles, en identifiant des thèmes, caractérisés par des associations de termes. Afin de compléter ces analyses, nous allons maintenant chercher à savoir si ces thèmes sont suffisamment représentatifs du contenu du corpus pour y classer les articles. Nous nous focaliserons ensuite sur les articles qui n'auront pas trouvé leur place dans ce regroupement afin d'en apprendre davantage sur les articles à la marge des travaux courants.

Classification de textes : À la suite de cette classification, 80% des articles ont été répartis dans onze classes disjointes (voir figure 4). Nous avons ensuite réalisé une deuxième classification sur le sous-corpus constitué des 3 160 articles non classés lors de la première analyse. Nous identifions cinq classes constituées de 99% des articles (voir figure 5). Seulement trois articles n'ont pu être classés.

À la suite de ces deux classifications, nous pouvons répondre à la question **Q2** en listant les thèmes étudiés dans les articles concernant les trajectoires de patients. Le premier thème est la maladie avec, par exemple, les troubles métaboliques comme le diabète et les complications cardiovasculaires. Certains articles concernent le ressenti du patient, ses angoisses et le vécu de sa maladie. Dans le parcours du patient, il y a le soutien par son environnement proche, la famille mais aussi les dispositifs mis en place comme l'intervention d'une infirmière à domicile. D'autres articles traitent de la fin de vie, des soins palliatifs et des procédés mis en place pour gérer cette dernière étape de la maladie. Un autre thème évoqué est la recherche clinique, avec la constitution de cohortes, la collecte de données, les méthodes employées dans ces différentes études. Ensuite, il y a l'organisation de l'hôpital, ses différents services, le personnel dont il dispose pour soigner les patients et les coûts associés. D'autres articles sont axés sur les réglementations et recommandations sanitaires régies par les guides de bonnes pratiques.

Avec cette méthode, il est relativement simple de repérer les thèmes qui nous intéressent plus spécifiquement, puis, soit de les explorer en pratiquant de nouvelles analyses de fouille textuelles, soit de les analyser avec des revues systématiques. Il est également aisé d'écarter tous les articles hors sujet, comme ceux traitant de la génétique, dont la thématique avait déjà été repérée par l'ADS. En d'autres termes, ces trois représentations nous donnent une vue d'ensemble des thèmes abordés dans ce corpus de textes, en organisant les articles dans ces catégories, ce qui permet à l'expert d'effectuer un premier tri en sélectionnant uniquement les articles d'intérêt pour son étude.

5 Discussions et Conclusions

Notre stratégie, sous forme de raffinements successifs, a permis de mettre en exergue les mots importants sous forme de nuages, puis de les connecter à d'autres mots avec l'ADS, mettant ainsi, en évidence un univers lexical. La classification des articles sans *a priori*, met en évidence les divers thèmes qui recouvrent l'ensemble de ces articles.

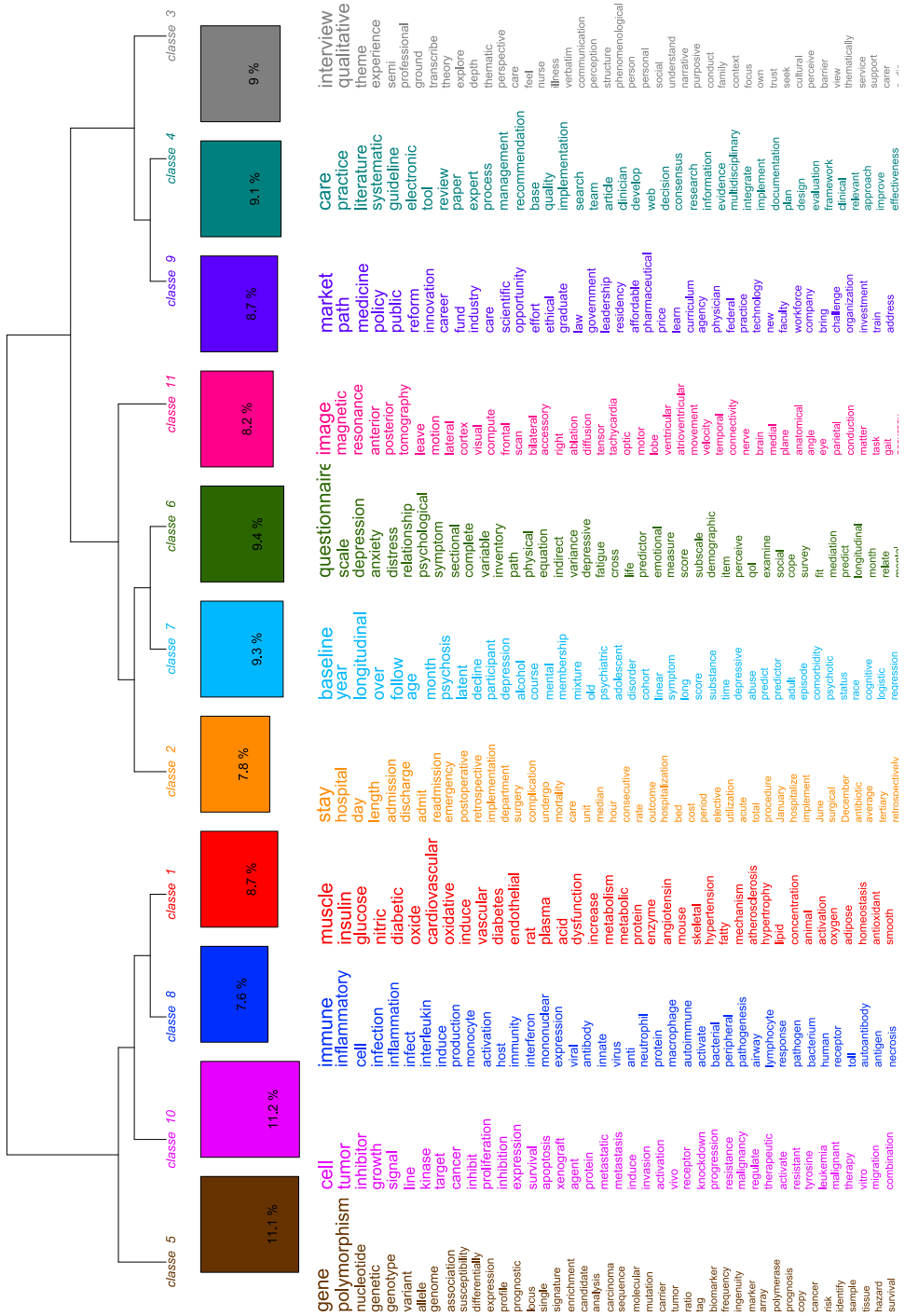


FIGURE 4 – Résultats d'une classification des articles pour le corpus Trajectoire

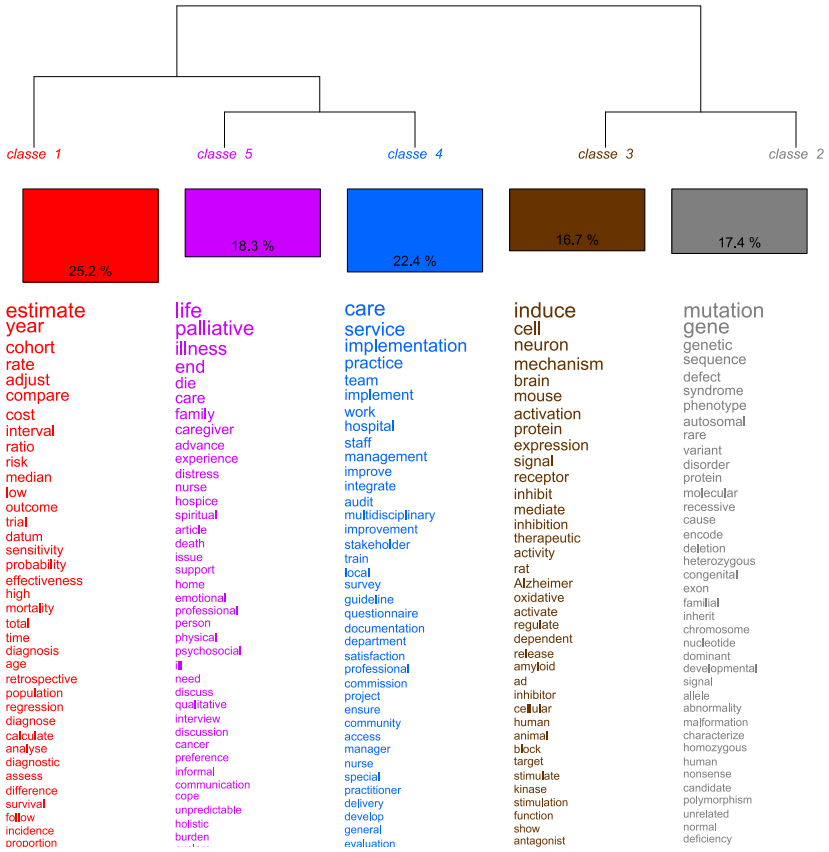


FIGURE 5 – Résultats de la deuxième classification des articles pour le corpus Trajectoire

L'ADS donne le contexte d'utilisation des mots. Cet outil d'analyse macroscopique offre une visualisation condensée et compartimentée des différents thèmes abordés par l'ensemble des articles. Dans cette étape, les articles sont considérés comme un ensemble à part entière. L'agencement des réseaux de mots permet de cerner l'importance d'un thème, de le placer par rapport aux autres dans un contexte déterminé, comme nous avons pu le faire avec le thème cancer. Cette représentation sous forme d'arbre, est particulièrement intéressante pour explorer non pas des articles mais un thème, sans se soucier de savoir s'il recouvre majoritairement un article.

La troisième étape de classification permet d'avoir un point de vue microscopique sur le corpus, celui des articles et de savoir si les thèmes sont représentatifs de ces articles. Ces derniers sont considérés comme des entités à part entière. Grâce à cette représentation, nous pouvons également mettre en évidence des thèmes qui recouvrent le corpus, mais aussi sélectionner les articles qui vont le plus nous intéresser, et *a contrario* identifier et retirer rapidement les documents hors sujet.

Cette approche a été efficace pour explorer le thème "Trajectoire" en collaboration avec un expert de la thématique qui a pu répondre à des questions sans *a priori* pour lesquelles il n'était pas possible de chercher une liste finie d'indicateurs dans les textes.

La recherche documentaire sur les trajectoires de patients, a montré que ce type d'étude

suscite un intérêt croissant dans la communauté biomédicale : que ce soit pour en apprendre davantage sur l'évolution d'une pathologie grâce au suivi du patient ou pour comparer les parcours de soins afin de mettre en place des stratégies par des procédures de soins facilitant le travail des personnels de santé tout en fournissant un cadre rassurant au patient et en réduisant les coûts. Nous retenons de cette étude que le concept de trajectoire est exploré plus particulièrement en cancérologie.

Cette première approche ouvre des perspectives intéressantes d'aide à l'analyse documentaire. À court terme, nous allons utiliser la méthode décrite dans cet article pour une analyse plus complète des trajectoires de patients, en croisant cette thématique avec le contexte d'étude des bases hospitalières de la tarification à l'activité mais également celui de l'infarctus du myocarde. Nous compléterons cette analyse avec une revue systématique en appliquant la méthode PRISMA. Nous utiliserons la grille de lecture associée à cette méthode et intégrerons de nouveaux indicateurs issus de l'exploration *a priori*, comme la définition du concept de trajectoire. À plus long terme, nous souhaitons améliorer la méthode proposée qui reste préliminaire. D'autres méthodes de fouille de textes peuvent être appliquées pour améliorer l'analyse bibliographique. Nous envisageons notamment la visualisation des termes d'intérêt correspondant aux items de la méthode PRISMA dans les publications pour aider l'analyse manuelle systématique (Abi-Haidar *et al.*, 2016) ou encore le résumé des thématiques qui consiste à prendre les points essentiels d'un texte pour en faire un paragraphe (Liu *et al.*, 2015).

Références

ABI-HAIDAR A., YANG B. & GANASCIA J.-G. (2016). Mapping the first world war using interactive streamgraphs. *Sociology and Anthropology*, **4**, 12–16.

BADA M. (2014). Mapping of biomedical text to concepts of lexicons, terminologies, and ontologies. *Methods in Molecular Biology (Clifton, N.J.)*, **1159**, 33–45.

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, **3**, pp. 993–1022.

BOLLEGALA D., OKAZAKI N. & ISHIZUKA M. (2010). A bottom-up approach to sentence ordering for multi-document summarization. *Information Processing and Management*, **46**(1), 89–109.

BRANDES U. (2001). A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology*, **25**(2), 163–177.

COHEN A. M., HERSH W. R., PETERSON K. & YEN P.-Y. (2006). Reducing Workload in Systematic Review Preparation Using Automated Citation Classification. *Journal of the American Medical Informatics Association*, **13**(2), 206–219.

CSARDI M. G. (2015). Package 'igraph'. *The Comprehensive R Archive Network. See http ://cran. r-project. org/web/packages/igraph/igraph. pdf.*

FLAMENT C. (1981). Similarity analysis : A technique for researches in social representations. *Cahiers de Psychologie Cognitive*, **1**(4), 375–395.

FLEUREN W. W. & ALKEMA W. (2015). Application of text mining in the biomedical domain. *Methods*, **74**, 97–106.

FRANTZI K., ANANIADOU S. & MIMA H. (2000). Automatic Recognition of Multi-Word Terms : the C-value/NC-value method. *International Journal on Digital Libraries*, **3**(2), 115–130.

FRUNZA O., INKPEN D., MATWIN S., KLEMENT W. & O'BLENIS P. (2011). Exploiting the systematic review protocol for classification of medical abstracts. *Artificial Intelligence in Medicine*, **51**(1), 17–25.

GEIFMAN N., BHATTACHARYA S. & BUTTE A. J. (2015). Immune modulators in disease : integrating knowledge from the biomedical literature and gene expression. *Journal of the American Medical Informatics Association*.

HIGGINS J. P., GREEN S. *et al.* (2008). *Cochrane handbook for systematic reviews of interventions*, volume 5. Wiley Online Library.

HUANG C.-C. & LU Z. (2016). Community challenges in biomedical text mining over 10 years : success, failure and the future. *Briefings in Bioinformatics.*, **17**(1), 132–44.

JOACHIMS T. (1998). Text Categorization with Support Vector Machines : Learning with Many Relevant. *Machine Learning : ECML-98*, **1398**, 137–142.

KELLY L. & ST PIERRE-HANSEN N. (2008). So many databases, such little clarity : Searching the literature for the topic aboriginal. *Canadian family physician Médecin de famille canadien*, **54**(11), 1572–1573.

LEITNER F. & VALENCIA A. (2008). A text-mining perspective on the requirements for electronically annotated abstracts. *FEBS Letters*, **582**(8), 1178–1181.

LIN J. M., BOHLAND J. W., ANDREWS P., BURNS G. A., ALLEN C. B. & MITRA P. P. (2008). An analysis of the abstracts presented at the annual meetings of the Society for Neuroscience from 2001 to 2006. *PLoS ONE*, **3**(4).

LIU F., FLANIGAN J., THOMSON S., SADEH N. & SMITH N. A. (2015). Toward abstractive summarization using semantic representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1077–1086.

MO Y., KONTONATSIOS G. & ANANIADOU S. (2015). Supporting systematic reviews using LDA-based document representations. *Systematic Reviews*, **4**.

MOHER D., LIBERATI A., TETZLAFF J. & ALTMAN D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses : the prisma statement. *British Medical Journal*, **339**.

RATINAUD P. & DÉJEAN F. (2009). IRaMuTeQ : implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre ». In *MASHS2009*, p. 1–22.

RATINAUD P. & MARCHAND P. (2012). Application de la méthode ALCESTE à de « gros » corpus et stabilité des « mondes lexicaux » : analyse du « CableGate » avec IRaMuTeQ ». In *Actes des 11ème Journées internationales d'Analyse statistique des Données Textuelles*, p. 835–844.

REINERT A. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, **VIII, (2)**, 187–198.

SEBASTIANI F. (2002). Machine Learning in Automated Text Categorization. *ACM Comput. Surv.*, **34**(1), 1–47.

SIEVERT C. & SHIRLEY K. (2014). LDAvis : A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, p. 63–70, Baltimore, Maryland, USA : Association for Computational Linguistics.

SONG M. (2014). Takes : Two-step Approach for Knowledge Extraction in Biomedical Digital Libraries. *Journal of Information Science Theory and Practice*, **2**(1), 6–21.

THOMAS J., MCNAUGHT J. & ANANIADOU S. (2011). Applications of text mining within systematic reviews. *Research Synthesis Methods*, **2**(1), 1–14.

VAZQUEZ M., KRALLINGER M., LEITNER F. & VALENCIA A. (2011). Text Mining for Drugs and Chemical Compounds : Methods, Tools and Applications. *Molecular. Informatics.*, **30**(Issue 6-7), 506–519.

Construction de ressources sémantiques pour l'amélioration de la qualité du clustering de messages courts

Yahaya Alassan Mahaman Sanoussi[1,2]

[1] MODYCO, Université Paris Nanterre, France
sanoussialassane@yahoo.fr

[2] Succeed Together, Paris, France
sanoussi@succeed-together.eu

Resumé

Prendre en compte l'aspect sémantique des données textuelles lors de la tâche de classification s'est imposé comme un réel défi ces dix dernières années. Cette difficulté vient s'ajouter au fait que la plupart des données disponibles sur les réseaux sociaux sont des textes courts, ce qui a notamment pour conséquence de rendre les méthodes basées sur la représentation "bag of words" peu efficientes. La plupart des approches présentes dans la littérature utilisent des connaissances externes comme wikipedia afin d'enrichir les messages courts avant la tâche de classification. Dans cet article, nous proposons la création de ressources permettant d'enrichir les messages courts afin d'améliorer la performance des méthodes de classification non supervisée. Pour constituer ces ressources, nous utilisons des techniques de fouille de données séquentielles.

Mots clés : **classification ; motifs fréquents ; motifs émergents ; ressources sémantiques**

1 INTRODUCTION

Le clustering (ou classification non supervisée) de messages courts est l'une des tâches les plus en vue durant cette dernière décennie dans le domaine du traitement automatique des langues. Cela est dû à la prolifération des données textuelles sur le web et les réseaux sociaux. L'exploitation de ces données est de plus en plus importantes pour les entreprises ; elle permet par exemple la mise en place de stratégies concurrentielles et marketing.

De nombreux chercheurs ont travaillé sur les tâches de classification non supervisée ; dans (Pavel Berkhin, 2002 ; Xu et Wunsch, 2005), les auteurs dressent un état de l'art détaillé de ces méthodes. Les conclusions qui en découlent montrent que les méthodes dépendent du type de données utilisées. Les méthodes ont fourni des résultats satisfaisants pour plusieurs applications sur des textes longs. Pour les textes courts, les résultats sont nettement moins bons. Leurs tailles limitées, l'usage important d'abréviations, l'utilisation d'acronymes ainsi que les problèmes liés au phénomène de la synonymie et de la polysémie apportent de nouveaux défis pour des applications d'exploration de données telles que le clustering de messages courts. Le modèle de représentation "bag of words" utilisé par la tâche du clustering est peu satisfaisant car les relations entre les termes ne sont pas exploitées. Plusieurs solutions sont présentes dans la littérature. Certains travaux préconisent des nouveaux modèles de représentation pour les messages courts ou l'enrichissement des messages en utilisant des ressources externes (Yang *et al.*, 2014). Dans (Hotho *et al.*, 2003), les auteurs montrent que l'utilisation des ressources externes sous forme d'ontologie peut améliorer la qualité du clustering de messages.
Dans cet article nous proposons de nous appuyer sur la construction de ressources sémantiques, en utilisant des techniques de fouille de données séquentielles pour extraire des motifs émergents. Ces derniers ont déjà été utilisés dans (Quiniou *et al.*, 2012) pour caractériser les genres de textes. Ainsi

nous extrayons des motifs émergents pour enrichir les messages courts dans le cadre de la classification non supervisée (clustering). Nous montrons que le clustering est amélioré grâce à cet enrichissement. Nous présentons l'approche pour la création et l'utilisation de ressources ainsi que les premiers résultats obtenus.

2 Construction de la ressource sémantique et enrichissement de messages courts

Dans cette section nous décrivons les données utilisées pour l'extraction des motifs puis nous présentons le processus de decouverte de motifs et enfin le processus d'enrichissement de messages courts.

2.1 Les données

Les données utilisées pour la construction de la ressource proviennent d'une enquête réalisée en novembre 2015 au sein d'une entreprise du secteur bancaire. Au cours de cette enquête, les employés du service informatique donnent leurs commentaires par rapport à l'appréciation globale de leur système d'information. C'est une enquête qui se répète, la même question est posée aux mêmes participants. Ces commentaires sont répartis dans 13 catégories différentes. Ces dernières sont les résultats d'un traitement semi-automatique car produites en utilisant un clustering automatique et validées par une équipe de consultants. Le tableau ci-dessous donne des informations sur la répartition des commentaires par catégorie.

Répartition des commentaires par catégorie		
Catégorie	**Thèmes de la catégorie**	**Nombre de commentaires**
Catégorie 1	bug, dysfonctionnement, problème technique	242
Catégorie 2	ergonomie, manque de fluidité, pas intuitif	183
Catégorie 3	lenteur, temps trop long, perte du temps	163
Catégorie 4	base documentaire et moteur de recherche inefficace	210
Catégorie 5	trop d'outil différents, trop de liens	125
Catégorie 6	complexité	90
Catégorie 7	trop de mots de passe, trop de codes d'accès	85
Catégorie 8	outils obsolètes	34
Catégorie 9	résolution d'incident lente	19
Catégorie 10	pas adapté	17
Catégorie 11	manque de formation	13
Catégorie 12	écran trop petit	11
Catégorie 13	problème d'imprimante	9

TABLE 1 – Les données utilisées pour la création de ressource

2.2 Extraction des motifs

N catégories sont proposées au processus d'extraction de motifs, chaque catégorie représente un ensemble de commentaires sémantiquement proches. Chaque commentaire est tout d'abord pré-traité (lemmatisation et l'utilisation d'une liste de stopwords). Les motifs séquentiels (fréquents) sont extraits pour chacune des catégories. Ensuite des motifs émergents sont calculés à partir des N collections de motifs fréquents extraits. La figure 1 illustre les différentes étapes du processus d'extraction de ressource mise en place.

2.2.1 Extraction des motifs fréquents

La fouille des motifs séquentiels est une technique de fouille de données dont l'objectif est d'extraire des connaissances sous forme de motifs (ou régularités) dans des bases de données dans les-

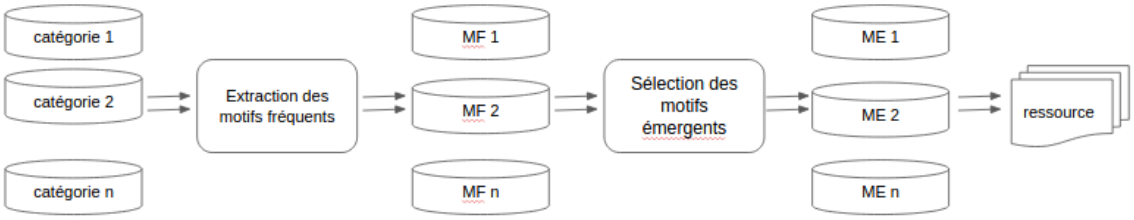


FIGURE 1 – Vue générale du processus de création de ressource

quelles l'ordre temporel caractérise les données (Agrawal et Srikant, 1995).

Soit $M = \{m_1, m_2, ..., m_n\}$ un ensemble de n attributs appelés des items. Dans le contexte de cet article les items sont des mots. Une séquence est une liste ordonnée d'items et est présentée par $S = \{i_1, i_2, ..., i_n\}$. La séquence dans notre cas est considérée comme un commentaire d'une personne par rapport à une question précise.

Base de données SDB	
1	former, collaborateurs,méthodes,ventes
2	former, équipiers
3	adaptation,collaborateurs,systèmes
4	former,accompagner,équipiers

TABLE 2 – base des données SDB de séquences

Le support d'une séquence S_1 dans une base de séquences SDB, noté, $sup(S_1)$ est le nombre des tuples contenant S_1 dans SDB. Par exemple dans le tableau précédant, le $sup\,(former, equipier)$ est égal à 2 car présent dans les séquences 2 et 4. Un motif fréquent est une séquence dont le support est supérieur à un seuil fixé. L'outil utilisé pour réaliser cette tâche est l'outil SDMC – Sequential Data Mining under Constraints – (Béchet *et al.*, 2015). Cet outil permet l'utilisation de contraintes afin d'extraire des motifs pertinents :

— La contrainte support minimal, au moins 2 dans notre cas : le support minimal est le nombre minimal de phrases dans lequel ce motif occurre. Cette contrainte traduit une certaine régularité des motifs produits.

— La contrainte de gap, au maximun 1 dans notre cas : Un motif séquentiel avec contrainte de $gap[M, N]$, noté $P[M, N]$ est un motif tel qu'au minimum M items et au maximum N items sont présents entre chaque item voisin du motif dans les séquences à partir desquelles il est extrait.

— La contrainte de longueur, au maximun 2 dans notre cas : pour ne concerver que les motifs de maximum 2 mots.

2.2.2 Sélection des motifs émergents

Initialement introduits dans Agrawal & Srikant (1995), les motifs émergents permettent de caractériser une classe d'objets par rapport aux autres classes. En effet, ils représentent les caractéristiques fortement présentes dans une classe et rares dans les autres. Dans (Quiniou *et al.*, 2012), un motif M d'un ensemble G_1 par rapport à un autre ensemble G_2 est émergent si $TauxCrioss(P) \geq \rho$ où

$$TauxCrioss(P) = \begin{cases} \infty & si\ supG_2(P) = 0 \\ \frac{sup_{G_1}(P)}{sup_{G_2}(P)} & sinon \end{cases}$$

$sup_{G_1}(P)$ et $sup\_G_2(P)$ désignent respectivement le support relatif du motif P par rapport à G_1 et celui par rapport à l'union des autres ensembles noté G_2.

La figure ci-dessous représente un extrait de la ressource sémantique extraite :

FIGURE 2 – Extrait de la ressource extraite

Le tableau suivant montre le nombre de motifs avant et après l'extraction des motifs émergents pour chacune des catégories :

Répartition des motifs par catégories		
Catégorie	**Motifs fréquents**	**Motifs emergents**
bug, dysfonctionnement, problème technique	362	282
ergonomie, manque de fluidité, pas intuitif	317	219
lenteur, temps trop long, perte du temps	294	209
base documentaire et moteur de recherche inefficace	434	361
trop d'outils différents, trop de liens	212	128
complexité	152	72
trop de mots de passe, trop de codes d'accès	200	130
outils obsolètes	38	15
résolution d'incidents lente	35	18
pas adapté	35	15
manque de formation	17	8
écran trop petit	6	2
problème d'imprimante	4	2

TABLE 3 – Motifs séquentiels avant et après l'extraction des motifs émergents

2.3 Enrichissement de messages courts

L'enrichissement consiste à utiliser la ressource comme étant un vecteur de champs sémantique. Pour un terme d'un message court (pré traité de la même manière que les données utilisées pour la découverte des motifs), on regarde s'il est présent dans la ressource afin d'ajouter (1 ou plusieurs fois) le nom de la thématique qui lui est associée.

Soit une ressource sémantique contenant la thématique **ergonomie** composée par les motifs suivants : `pas fluide; pas ergonomique;manque convivialité; intuitif`. À partir de deux messages courts suivants :`mon système manque de convivialité` et `logiciel pas fluide`, ce processus permet de les enrichir et d'obtenir : `mon système manque de convivialité` **`ergonomie`** et `logiciel pas fluide` **`ergonomie`**. Cela permet de mettre en évidence la sémantique des messages courts en unifiant le vocabulaire utilisé. En effet les deux messages courts ne partageaient au départ pas de mots en commun, ce qui rendait difficile leur classification dans une même catégorie. Grâce à l'enrichissement, un terme en commun apparait dans les deux messages, ce qui facilitera leur catégorisation.

3 Évaluation

Afin d'évaluer l'impact de la ressource sémantique (cf 2) sur le regroupement de messages courts, nous utilisons la méthode du clustering de ward – Ward's Hierarchical Clustering Method : Clustering Criterion and Agglomerative Algorithm – (Murtagh & Legendre, 2011). Il s'agit de produire un regroupement sur les données de reference (avec et sans enrichissement) et après comparer les groupes prédits et les groupes de reference en utilisant les mesures décrites dans la section 3.2.

3.1 Données de reference

Les données de reference proviennent d'une enquête réalisée en décembre 2015, similaire à celle qui a servie pour la création de ressource sémantique (cf 2.2). Les employés du service informatique d'un des groupes de l'entreprise ont donné leurs appréciation par rapport au système d'information qu'ils utilisent. Ces données ont été pré-traitées et catégorisées par un algorithme de clustering et validées par une équipe de consultants. Ces données constituent les données de reference. Le tableau ci-dessous donne des informations sur la repartition des commentaires par catégories.

Répartition des commentaires par catégories		
Catégorie	**Thèmes de la catégorie**	**Nombre de commentaires**
Catégorie 1	manque d'intuitivité	40
Catégorie 2	revoir et actualiser la base documentaire	25
Catégorie 3	lenteur du système en général	24
Catégorie 4	dysfonctionnement regulier	21
Catégorie 6	moteur de recherche peu efficace	19
Catégorie 7	difficulté à trouver les bons éléments	14
Catégorie 8	trop d'outils différents	12
Catégorie 9	identification et mots de passe disparates	6
Catégorie 10	système d'information archaique et précaire	5
Catégorie 11	environnement pas adapté	4
Catégorie 12	aucune formation	3
Catégorie 12	certains outils sont bien mais ...	1

3.2 Mesures d'évaluation

Deux mesures de qualité du clustering sont utilisées (Rosenberg & Hirschberg, 2007) :

— Homogeneity : Seuls les messages courts d'un même groupe des données de références doivent être assignés dans une même classe par un algorithme de clustering.

— Completeness : Les messages courts d'un même groupe de données de références doivent être toujours assignés dans une même classe par un algorithme de clustering.

3.3 Résultats

D'une part, il s'agit ici de voir l'évolution du coefficient homogeneity et completness en fonction du nombre d'ajout de l'information (nom des thématiques ajoutés de 1 à 10 fois).D'autre part, nous pouvons comparer le clustering avec et sans enrichissement. La figure 3 nous montre que le clustering avec enrichissement (courbe rouge) donne des meilleurs résultats par rapport au clustering sans enrichissement (courbe verte). Une amélioration (en ajoutant le nom des thématiques 8 fois) de 42% (0,54 versus 0,76) pour la completeness et 45% (0,58 versus 0,84) pour homogeneity.

FIGURE 3 – Evaluation clustering

4 Conclusion

L'objectif de cet article est de montrer à quel point notre approche de construction et d'utilisation de la ressource peut contribuer à améliorer la performance des applications dédiées au regroupement des messages courts. Les résultats sont prometteurs. Dans les travaux futurs, il serait intéressant de faire varier les paramètres utilisés lors de l'extraction de la ressource (par exemple le seuil servant à sélectionner les motifs emergents). Enfin, il serait important d'appliquer cette méthode sur plusieurs jeux de données et secteurs d'activités différents pour la valider totalement.

Références

AGRAWAL R. & SRIKANT R. (1995). Mining sequential patterns. In *ICDE'95*.

BÉCHET N., CELLIER P., CHARNOIS T. & CRÉMILLEUX B. (2015). Sequence mining under multiple constraints. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, p. 908–914.

HOTHO A., STAAB S. & STRUMME G. (2003). Ontologies improve text clustering. In *ICDM*.

MURTAGH F. & LEGENDRE P. (2011). Ward's hierarchical clustering method : Clustering criterion and agglomerative algorithm. In *Retrieved from http ://arxiv.org/pdf/1111.6285.pdf*.

QUINIOU S., CELLIER P., CHARNOIS T. & LEGALLOIS D. (2012). What about sequential data mining techniques to identify linguistic patterns for stylistics ? In *Proc. of CICLing'2012*, New Delhi, India.

ROSENBERG A. & HIRSCHBERG J. (2007). V-measure : A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL*, p. 410–420.

YANG C.-L., BENJAMASUTIN N. & CHEN-BURGER Y.-H. (2014). Mining hidden concepts : Using short text clustering and wikipedia knowledge. In *WAINA14*, p. 675–680.

Recherche d’information et Extraction de connaissances

L'apprentissage d'ordonnancement pour l'appariement de questions

Camille Pradel, Baptiste Chardon,
Dominique Laurent, Sophie Muller, Patrick Séguéla
Synapse Développement, 5 rue du Moulin-Bayard, 31000 Toulouse,
{ camille.pradel, baptiste.chardon, dlaurent, sophie.muller,
patrick.seguela }@synapse-fr.com

Résumé : Cet article présente une approche permettant à un utilisateur d'interroger une base de connaissances type FAQ, c'est-à-dire un ensemble de questions et leurs réponses respectives rédigées en langue naturelle. Le composant présenté dans cet article apparie la question de l'utilisateur à une ou plusieurs questions de la base de connaissances. Pour cela, nous utilisons un composant déjà existant d'analyse de questions, capable de sélectionner un ensemble de questions candidates proches de la question utilisateur, et de produire des traits propres à chaque couple (question utilisateur, question candidate). Ce composant est chaîné à un modèle permettant l'ordonnancement des questions candidates, qui est appris automatiquement de façon supervisée, une partie seulement du corpus d'apprentissage étant annotée manuellement, et le reste grâce à des règles ad-hoc. Ces travaux reprennent les résultats d'un domaine de recherche récent, l'apprentissage d'ordonnancement (*Learning to Rank*), et les adaptent à une application industrielle innovante, l'appariement de questions comme paradigme d'accès à la connaissance. Une expérimentation évalue sur des données issues d'un système en production la qualité de chacune des phases d'apprentissage.

Mots-clés : Interface en langue naturelle, FAQ, *Learning to rank*, apprentissage actif, recherche d'information.

1 Introduction

Une des tâches les plus populaires en Recherche d'Information (RI) consiste à retrouver un ensemble de documents pertinents vis-à-vis d'un besoin en information exprimé dans une requête. Pour cela, des approches efficaces existent depuis de nombreuses années ; elles sont fondées sur des heuristiques et n'exploitent généralement pas des modèles appris automatiquement à partir de corpus d'entrainement. Bien que simples et élégantes sur le papier, ces approches nécessitent pour être déployées en production une phase de réglage lourde et fastidieuse, tout particulièrement lorsque le nombre de paramètres à prendre en compte augmente (les moteurs de recherche du web prennent en compte plusieurs dizaines de paramètres et doivent sans cesse en intégrer de nouveaux pour rester compétitifs). L'idée d'apprendre automatiquement les modèles de classement prenant en compte tous ces paramètres devient de plus en plus séduisante au fur et à mesure que la mise en œuvre des algorithmes traditionnels se complexifie, d'autant plus que l'acquisition des données d'entrainement est, dans ce cas, simple et peu onéreuse (le choix de l'utilisateur parmi les résultats qui lui sont retournés étant un très bon indice de la pertinence de ces résultats). Yahoo ! et Microsoft ont successivement libéré des jeux de données et organisé des compétitions pour stimuler les recherches dans ce domaine.

Dans cet article, nous voulons exploiter ces mêmes méthodes et les appliquer à un domaine différent, celui de l'appariement de questions. L'objectif du système présenté dans la suite est d'apparier une question exprimée par un utilisateur à une ou plusieurs questions cibles issues d'une base de connaissances. Dans ce contexte, nous appelons base de connaissances un ensemble de questions cibles et leurs réponses respectives rédigées en

langue naturelle, une même réponse pouvant éventuellement être associée à plusieurs formulations de la même question, permettant ainsi de surmonter plus facilement la variabilité de la langue. Cet appariement peut notamment être utilisé pour suggérer les questions cibles dans un champ de recherche avec autocomplétion, ou bien au sein d'un agent conversationnel. Plus généralement, l'appariement de questions constitue une alternative aux questions-réponses particulièrement pertinente pour l'accès au savoir en milieu industriel, dont les avantages sont discutés en section 2.

Le système développé exploite le composant d'analyse de requête du moteur de question-réponse de Synapse développement. Ce composant est capable de sélectionner un ensemble de questions de la base de connaissances, appelées questions candidates dans la suite, qui sont sémantiquement proches de la question utilisateur. Il extrait aussi certaines caractéristiques de ces questions, comme leur type (recherche d'une date de naissance ou de mort, d'une durée, d'une ville, d'une personnalité…), leur objet (ce sur quoi elles portent) et les mots-clés qui les composent. Ce composant produit également un score pour chacune des questions appariées. De ce score, on pourrait immédiatement déduire un ordonnancement. Cependant, le système d'origine ayant pour objectif de chercher dans un texte la réponse à une question, ce score, bien que significatif, ne prend pas en compte certains des aspects essentiels à l'appariement de questions (par exemple, est-ce que les deux questions comparées sont de même type ?).

C'est pourquoi nous avons défini dans un premier temps un ensemble de règles permettant de (ré)ordonnancer les résultats produits par ce composant. Le système ainsi obtenu fonctionne suffisamment bien pour être utilisé en production. Mais, dans le but d'éviter d'avoir à maintenir un système de règles toujours plus complexe, nous voulons apprendre un modèle qui produit dans un premier temps des sorties similaires à ce système de règles puis qui saura se raffiner et s'adapter aux usages en exploitant les retours utilisateurs pour évoluer.

Nous justifions la pertinence de notre approche dans la section 2 et donnons un aperçu des méthodes existantes en apprentissage d'ordonnancement en section 3. Puis nous décrivons les travaux réalisés en section 4 et présentons en section 5 les résultats des évaluations. Enfin, nous concluons et présentons quelques perspectives en section 6.

2 Vers un nouveau paradigme d'accès à la connaissance

Dans cette section, nous justifions notre choix d'exploiter l'appariement de questions en langue naturelle plutôt que l'approche de questions-réponses traditionnelle par trois éléments : la difficulté moindre de la tâche d'appariement, le besoin de réassurance de l'utilisateur et le format des connaissances actuellement disponibles dans le milieu industriel.

2.1 Limites des interfaces entre langue naturelle et texte

Les interfaces entre la langue naturelle et les textes, ou systèmes de questions-réponses, ont fait l'objet de recherches très actives dans les années 2000. Ces approches exploitent une collection de documents comme source de connaissances et cherchent directement la ou les réponses à une question utilisateur dans ces documents. Le composant d'analyse de questions utilisé dans les expérimentations dans la suite de l'article est issu de *QRISTAL*, le système de question-réponse de *Synapse Développement* (Laurent & al., 2005).

D'après (Hirschman & Gaizauskas, 2001), le processus d'interprétation de la grande majorité de ces systèmes se divise en deux étapes :

1. identifier le type de l'objet de la requête ; une bonne reconnaissance du type de l'objet est primordiale, et de nombreux travaux se sont penchés sur le problème, en définissant des hiérarchies de types à partir du type de la réponse attendue (Moldovan & al., 1999; Hovy & al., 2000; Wu & al., 2003; Srihari & Li, 1999) ;
2. déterminer les contraintes additionnelles exprimées ; ces contraintes sont le plus souvent issues des relations syntaxiques et sémantiques entre les termes de la

requête en langue naturelle (Moldovan & al., 2002; Litkowski, 2000; Attardi & al., 2002).

En comparaison aux interfaces entre langue naturelle et bases de données qui les ont précédés, les systèmes de question-réponse sont complexes, du fait de la non-restriction du domaine et de la forme beaucoup moins contrainte et plus variable de la source de connaissances. Ils intègrent de nombreuses sous-tâches, parmi lesquelles on peut citer la reconnaissance d'entités nommées, l'extraction de relations, la résolution de coréférences et la désambiguïsation.

Un sous-ensemble de ces systèmes, apparu un peu plus tard, considère le web dans son intégralité comme corpus de documents. La tâche reste la même mais certains problèmes, comme le passage à l'échelle et l'hétérogénéité des données, prennent toute leur importance. Pour faire face à l'immensité du web, la majorité des systèmes proposés dans la littérature exploite des moteurs de recherche « classiques » par mots-clés, comme Google. La requête utilisateur est d'abord transformée en une ou plusieurs requêtes compatibles avec le moteur de recherche choisi, puis ces requêtes sont exécutées sur le web, et enfin les réponses sont extraites des documents les plus pertinents renvoyés par le moteur de recherche.

Les recherches visant à la résolution de cette tâche ont été largement orientées, de 1999 à 2007, par la compétition question-réponse sans restriction de domaine (*open domain question answering*) de la conférence sur l'extraction de données dans les textes (*TREC - Text REtrieval Conference*). Les systèmes développés durant ces années ont obtenu d'excellents résultats. Les recherches dans le domaine sont beaucoup moins actives depuis quelques années, et on tend à considérer ce problème comme résolu.

On constate cependant que la mise en place réelle de tels systèmes auprès d'utilisateurs finals n'a jamais été réalisée, si ce n'est pour des besoins de démonstration. Leur développement, notamment à l'échelle du web, se heurte à des verrous de taille (passage à l'échelle, gestion du bruit, identification des doublons...), et les solutions trouvées nécessitent d'importants efforts d'implémentation. Il nous semble que les systèmes proposés au cours des compétitions TREC souffrent de suradaptation aux données des compétitions et que leur évolution dans un contexte industriel n'est pas triviale.

2.2 Utilisabilité des interfaces en langue naturelle

La pertinence des interfaces en langue naturelle a été mise en cause plusieurs fois dans la littérature, sans jamais pour autant mener à une conclusion claire (Chakrabarti, 2004; Dekleva, 1994; Desert, 1993; Thompson & al., 2005).

Plusieurs études (Dittenbach & al., 2003; Reichert & al., 2005) ont voulu analyser l'intérêt des utilisateurs finals pour les requêtes en langue naturelle, par rapport aux requêtes par mots-clés. Les auteurs de (Dittenbach & al., 2003) ont ainsi développé une interface test de requêtes libres dans le domaine du tourisme, et récolté 1425 requêtes spontanées d'utilisateurs. 57,05% des requêtes étaient des phrases complètes exprimées en langue naturelle et grammaticalement correctes ; 21,26% étaient des fragments de questions, du type « double room for two nights in Vienna » ; 21,69% étaient des requêtes mots-clés. D'après les auteurs, les interfaces en langue naturelle se montrent particulièrement utiles lorsque le public visé est hétérogène, comme dans le cas de la plate-forme utilisée. Les auteurs de (Reichert & al., 2005) ont demandé à des étudiants d'utiliser deux versions différentes d'un outil de question-réponse pour le *e-learning*, *CHESt* (Linckels & Meinel, 2005). Dans une version, les requêtes s'expriment sous forme de mots-clés ; dans l'autre, elles s'expriment en langue naturelle. 76% des étudiants ont déclaré préférer la version mots-clés mais concèdent également qu'ils utiliseraient plus volontiers la version langue naturelle si ils savaient pouvoir obtenir de meilleurs résultats.

Dans (Kaufmann & Bernstein, 2010), les auteurs évaluent l'utilisabilité des interfaces permettant aux utilisateurs d'accéder à des connaissances par le biais de la langue naturelle. Ils identifient trois principaux problèmes liés à ce type d'interface :

- l'*ambiguïté* de la langue naturelle est le plus évident, l'*ambiguïté linguistique* ─ une même expression en langue naturelle peut être interprétée de différentes façons ─ va de pair avec la *variabilité linguistique* ─ une même idée peut s'exprimer de différentes façons en langue naturelle.
- la *barrière adaptative* (*adaptivity barrier*) rend les interfaces présentant de bonnes performances de recherche peu portables ; ces systèmes sont la plupart du temps spécifiques au domaine et, par conséquent, difficiles à adapter à de nouveaux contextes. L'ensemble des connaissances ontologiques contenues dans une base de connaissances est vu comme un moyen réaliste de pallier ce problème.
- le *problème d'habitabilité* (*habitability problem*) selon lequel l'utilisateur final peut se sentir perdu face à une trop grande liberté dans l'expression de sa requête. La plupart des interfaces en langue naturelle n'expliquent pas comment exprimer ses requêtes ou quel type d'information peut être demandé ; en conséquence, l'utilisateur peut exprimer des requêtes au-delà des capacités du système (que ce soit du point de vue du besoin en information ou de l'expression de la requête) ou, pire encore, le système peut mal interpréter la requête et l'utilisateur peut ne pas s'en rendre compte et considérer une réponse inappropriée comme satisfaisante.

La contribution principale de (Kaufmann & Bernstein, 2010) est inspirée de ce dernier problème : l'*hypothèse d'habitabilité* prétend qu'une interface en langue naturelle, pour présenter la meilleure utilisabilité, devrait imposer une certaine structure à l'utilisateur dans le but de l'orienter lors du processus de formulation de la requête. Ainsi le « syndrome de la feuille blanche » devrait être évité. Cependant, la contrainte structurelle ne doit pas être trop restrictive, le risque étant d'aliéner l'utilisateur.

Pour soutenir cette hypothèse, les auteurs décrivent l'étude d'utilisabilité qu'ils ont conduite auprès d'utilisateurs. Quatre interfaces de requêtes ont été développées et confrontées à 48 utilisateurs. Les auteurs proposent de situer chaque interface sur un *continuum de formalisation* (*formality continuum*) en fonction de la nature de l'interaction avec l'utilisateur final. Les approches complètement LN sont à une extrémité de ce continuum, alors que les langages formels de requêtes sont à l'autre extrémité. Les résultats de cette étude montrent clairement que, conformément à l'hypothèse d'habitabilité, les utilisateurs se sentent le plus à l'aise en utilisant Querix et son langage de requêtes légèrement contraint.

Ce point de vue est renforcé par des travaux plus récents présentés dans (Damljanović et al., 2013) qui exploitent des retours utilisateurs et des fenêtres de clarification pour augmenter l'expérience de l'utilisateur.

Nous retenons de ces résultats le besoin d'un retour de l'interface vers l'utilisateur afin de rassurer ce dernier quant aux informations qui lui sont retournées en réponse à sa requête.

2.3 Bases de connaissances sous forme de FAQ

Il est possible de trouver chez des grands groupes industriels, notamment aéronautiques et pharmaceutiques, des bases de connaissances structurées, par exemple sous forme de graphe, avec des schémas cohérents et des données fiables respectant ces schémas. Ces ressources critiques ne sont généralement pas accessibles publiquement.

Les entreprises de taille plus modeste n'investissent en général pas dans la construction de telles ressources, très onéreuse, et la grande majorité des connaissances métiers des entreprises reste à ce jour encore exprimée sous forme textuelle.

Parmi ces ressources textuelles, une partie est rédigée sous forme de FAQ (pour *Frequently Asked Questions*), c'est-à-dire de couples (question, réponse) exprimés en langue naturelle. Ce format est en effet traditionnellement utilisé pour présenter des informations à l'utilisateur final d'un service. Il est censé permettre de trouver plus rapidement une réponse à un besoin précis en information ; le lecteur se contente de parcourir les questions jusqu'à trouver celle qui l'intéresse et évite ainsi une lecture exhaustive du document. Cette lecture reste cependant linéaire et, si la FAQ devient trop importante, une simple recherche peut devenir très laborieuse. Certaines entreprises, typiquement les éditeurs de logiciels, ont

construit des FAQ très importantes (souvent appelées bases de connaissances) ; elles peuvent être hiérarchisées par catégories de questions, mais, même ainsi, leur parcours manuel n'a plus de sens, car trop inefficace.

Il existe donc toujours un besoin de valoriser ces connaissances exprimées sous forme de FAQ et de faciliter leur accès par un utilisateur final, et c'est à ce besoin que l'approche que nous développons dans ce papier veut répondre.

2.4 L'appariement de questions

Pour les raisons exprimées dans les sous-sections précédentes, nous travaillons depuis peu sur l'évolution de notre moteur de question-réponse vers un nouveau mode d'interrogation des connaissances. Celui-ci est fondé sur l'appariement de questions d'une base de connaissances à la question de l'utilisateur.

Dans cette approche, l'expression du besoin en information de l'utilisateur peut par exemple se faire au travers d'un champ de recherche avec une autocomplétion suggérant des questions proches. L'utilisateur est ainsi confronté à un mode d'interaction qui lui est familier, il bénéficie d'un retour du système et porte ainsi une plus grande confiance aux informations qui lui sont retournées. S'il sélectionne une des questions suggérées, équivalente ou proche de sa question d'origine, il est certain que la réponse qui lui est affichée répond à la question sélectionnée. S'il renseigne sa propre question jusqu'au bout et la valide, c'est la réponse à la question la plus proche qui lui est soumise, mais l'interface devrait afficher la question appariée.

Selon l'hypothèse d'habitabilité de (Kaufmann & Bernstein, 2010), introduite ci-dessus, l'interface que nous proposons, située au milieu du continuum de formalité, offre un *feedback* qui permet d'augmenter la réassurance de l'utilisateur et de maximiser son expérience.

3 Learning to rank

L'apprentissage d'ordonnancement (*learning to rank*) est un domaine de recherche relativement récent ; l'objectif est de développer des algorithmes d'apprentissage pour des tâches d'ordonnancement.

Le principal domaine d'application est la RI (Liu, 2009), mais des travaux ont également été menés dans le traitement automatique des langues, pour l'ordonnancement de traductions potentielles d'une phrase en traduction automatique (Li, 2014). Une vue d'ensemble des principales approches existantes est donnée dans (Li, 2011).

Les travaux réalisés à ce jour sont en grande majorité issus de participations aux compétitions organisées par Yahoo ! et Microsoft, et ont été évalués sur les données fournies dans le cadre de ces compétitions. Les résultats actuels en apprentissage d'ordonnancement sont donc liées aux problématiques du domaine de la Recherche d'Informations (RI). Ces méthodes d'apprentissage ont déjà été expérimentées pour la recherche au sein de bases de connaissances type FAQ (Surdeanu & al., 2008 ; Bunescu & Huang, 2010)

3.1 Les catégories d'approches

L'ensemble des méthodes proposées jusque-là se divise en trois catégories. Les approches points par points (*pointwise approaches*) transforment le problème d'ordonnancement en un problème d'apprentissage automatique classique, comme la classification ou la régression ; il est ainsi possible d'exploiter les algorithmes de la littérature liés à ces problèmes. Les principales approches dans ce domaine comptent Subset Ranking (Cossock & Zhang, 2006), *McRank* (Li & al., 2007), *Prank* (Crammer & Singer, 2001), et *OC SVM* (Shashua & Levin, 2002).

Dans les approches par paires (*pairwise approaches*), l'ordonnancement est associé à un ensemble de classification par paires ou de régression par paires ; là encore, les algorithmes d'apprentissage existants peuvent être exploités. *Ranking SVM* (Herbrich & al., 1999), *RankBoost* (Freund & al., 2003), *RankNet* (Burges & al., 2005), *GBRank* (Zheng et al., 2008), *IR SVM* (Cao et al., 2006), et *LambdaMART* (Wu et al., 2010) font partie de ces approches.

La dernière catégorie de méthodes est spécifique à la tâche d'ordonnancement. Ces approches sont les seules à exploiter la structure de groupe ordonné des données d'apprentissage. C'est dans ce cas l'ordonnancement entier qui est considéré comme une instance d'apprentissage. Les principales approches de cette catégorie sont *ListNet* (Cao et al., 2007), *ListMLE* (Xia et al., 2008), *AdaRank* (Xu & Li, 2007), *SVM MAP* (Yue et al., 2007), et *Soft Rank* (Taylor et al., 2008).

3.2 Métriques d'évaluation

Tout comme en RI, les métriques d'évaluation représentent un aspect essentiel de l'apprentissage d'ordonnancement. Elles définissent la fonction objectif à optimiser lors des phases d'apprentissage et de validation. L'objectif de ces mesures est d'exprimer à quel point l'ordonnancement proposé par le système évalué est proche de l'ordonnancement de référence. Les métriques utilisées pour l'apprentissage d'ordonnancement sont à ce jour celles de la RI : Discounted Cumulative Gain (DCG) (Järvelin & Kekäläinen, 2002), Normalized Discounted Cumulative Gain (NDCG), Mean Average Precision (MAP), Kendall's Tau, Reciprocal Rank (RR) (Voorhees & Tice, 1999), Expected Reciprocal Rank (ERR) (Chapelle & al., 2009).

4 Construction du modèle d'ordonnancement

Comme expliqué plus haut, le modèle d'ordonnancement est construit en deux étapes :

1. Des règles simples prédéfinies sont utilisées pour produire des ordonnancements sur des questions appariées à un ensemble des questions utilisateurs issues d'un log de système en production permettant d'interroger une FAQ. On obtient ainsi un ensemble de couples (question utilisateur, ordonnancement de questions candidates), suffisamment important pour initier le modèle, qui intègre alors l'« intelligence » des règles définies précédemment.
2. En interagissant avec le système, les utilisateurs produisent des informations qui peuvent être exploitées pour raffiner le modèle. On considère en effet que lorsque l'utilisateur sélectionne une question dans la liste des suggestions qui lui sont proposées, cette question est la plus proche de son besoin en information initial, et on peut intégrer cette connaissance au modèle afin de favoriser la question sélectionnée lorsqu'une nouvelle question entrante présente les mêmes caractéristiques.

La figure 1 présente l'architecture générale du composant d'appariement de questions. Les deux cadres en pointillé représentent chacun une des phases d'apprentissage du modèle. Dans la phase *batch learning*, des ordonnancements issus de règles sont utilisés comme exemples. Puis dans la phase *active learning*, des informations sont recueillies auprès de l'utilisateur dans le but d'améliorer le modèle.

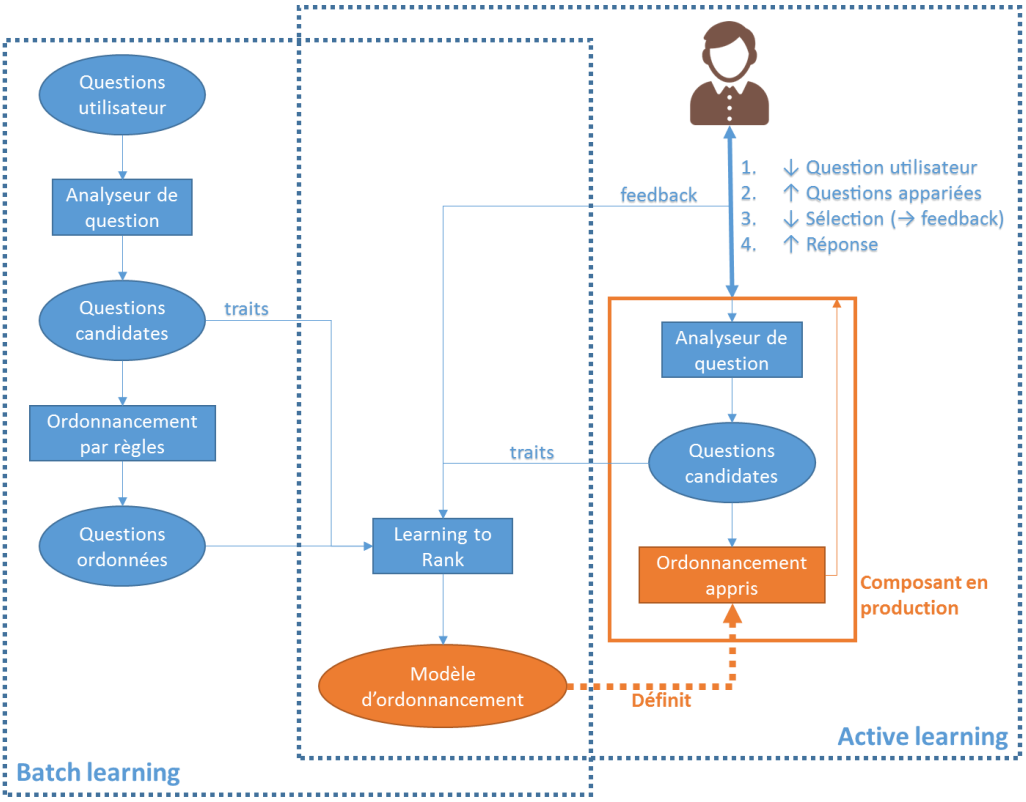


FIGURE 1 – Vue d'ensemble du système d'appariement de questions.

4.1 Génération du corpus d'apprentissage

Les logs que nous avons utilisés comportent 4855 requêtes exprimées par des utilisateurs. Il est important de préciser que ces 4855 ne sont pas toutes des questions. Dans certaines, le besoin en information est exprimé sous forme de mot-clés, comme par exemple « problème installation » ; d'autres n'expriment aucun besoin en information, on trouve par exemple des « bonjour », « merci », « au revoir », mais aussi des insultes ou d'autres entrées improbables, certainement dues à une mauvaise manipulation, par exemple « i ». Ces caractéristiques apportent une difficulté supplémentaire qui est abordée dans la section Perspectives.

Pour construire le corpus d'apprentissage, nous avons analysé chacune des 4855 requêtes afin d'obtenir, pour chaque requête, l'ensemble des questions candidates de la base de connaissances et les traits associés à chaque couple (question utilisateur, question candidate). Puis, sur ces 4855 ensembles de couples, 500 ont été ordonnés à la main pour simuler un retour utilisateur et les 4355 autres à l'aide du système de règles. Dans les expérimentations décrites plus bas, les trois quarts des données annotées à la main ont été utilisés pour l'apprentissage, le dernier quart a été conservé pour les tests.

Dans la littérature, pour exprimer le classement des requêtes candidates pour une même requête utilisateur, à chaque couple (question utilisateur, question candidate) est associée une note de pertinence, qui exprime à quel point le besoin exprimé dans la question candidate est proche de celui exprimé dans la question utilisateur. Cette note peut prendre une valeur entière de 0 à 4, 0 signifiant que les deux questions n'ont rien en commun, et 4 qu'elles sont extrêmement proches, voire identiques.

Nous avons utilisé cette même approche dans le but de simplifier la réexploitation des algorithmes existants. Lors de la génération des ordonnancements à partir des règles, la première question se voit attribuer la valeur 4, la deuxième 3, la troisième 2, la quatrième 1 et toutes les suivantes 0. Dans le cas où le composant d'analyse de requêtes ne retourne qu'une

seule requête candidate, nous lui attribuons la note 3 (cette décision a été prise après une observation qualitative d'un extrait des sorties de l'analyseur de question).

La constitution de l'ordonnancement manuel a été effectuée par un seul annotateur avec pour seule consigne de calquer la note utilisée dans les jeux de données publiés par Microsoft : « *The relevance judgments […] take 5 values from 0 (irrelevant) to 4 (perfectly relevant)* ».

4.2 Choix des traits

Pour constituer les données d'apprentissage, nous avons intégré l'ensemble des traits retournés par le composant d'analyse de question. Cela assure de donner à l'algorithme d'apprentissage tous les éléments pour intégrer l'intelligence codée dans les règles. Ces traits sont les suivants : 1. le score introduit plus haut qui représente la similarité entre la question utilisateur et la question candidate, 2. le nombre de mots-clés que la question candidate a en commun avec la question utilisateur, 3. le nombre de mots-clés distincts communs à la question candidate et à la question utilisateur, 4. un booléen indiquant si les deux questions sont de même type.

Nous avons également ajouté d'autres traits dans le but de permettre à l'algorithme d'apprendre des subtilités qui ne sont pas représentées dans les règles mais pourraient être déduites des feedbacks utilisateur :

- Quatre traits liés à la ressemblance entre les chaines de caractères de chaque question : 1. la distance de Levenstein entre les deux questions, 2. un booléen indiquant si les chaines de caractères sont identiques après neutralisation des accents et des majuscules, 3. un booléen indiquant si la chaine de la question utilisateur contient la chaine de la question candidate, 4. un booléen indiquant si la chaine de la question candidate contient la chaine de la question utilisateur,
- Deux traits visant à prendre en compte la « popularité » des questions de la base de connaissances, c'est-à-dire dans quelle mesure ces questions représentent les besoins en information généralement exprimés ; ces notes sont calculées sur l'ensemble des données d'entrainement au cours d'un pré-traitement : 1. un entier exprimant combien de fois une question apparaît dans une liste de questions candidates, 2. un entier exprimant combien de fois une question obtient la note de pertinence maximale.

4.3 Choix de la métrique

Pour choisir la métrique utilisée pour l'apprentissage (et pour l'évaluation présentée plus bas), nous considérons une intégration du système dans un champ de recherche avec une autocomplétion. Dans ce scénario, l'utilisateur veut se voir suggérer au plus vite une question qui correspond à son besoin en information. Il veut la réponse à sa question et rien de plus ; il n'est pas à la recherche de l'exhaustivité, comme lorsque l'on fait une recherche approfondie ou de la veille.

Nous avons pour objectif que l'utilisateur n'ait pas à consulter plus de trois suggestions avant de trouver celle correspondant à son besoin en information. Nous utilisons donc la mesure ERR@3 (Chapelle & al., 2009), qui représente la probabilité que l'utilisateur ait trouvé ce qu'il cherche dans les trois premières suggestions, tout en privilégiant les suggestions pertinentes tout en haut du classement.

4.4 Augmentation du corpus d'entrainement avec les questions de la base de connaissances

Le log de requêtes (et par conséquent le corpus d'entrainement) en notre possession étant de taille relativement réduite pour mettre en œuvre un apprentissage automatique, nous avons

voulu tester l'utilisation d'un artifice permettant de le gonfler légèrement avec des données fiables.

Nous avons pour cela augmenté notre corpus avec l'ensemble des questions de la base de connaissances, que l'on considère alors comme des questions utilisateurs. En sortie du système de règles, chacune de ces questions est logiquement associée à elle-même avec le score maximal, nous voulons ainsi permettre à l'algorithme d'apprentissage de prendre en compte les traits liés aux chaines de caractères décrits plus haut.

4.5 Augmentation de l'influence des données annotées manuellement

Les données annotées à la main sont présentes en moins grande quantité dans le corpus d'apprentissage et influencent donc beaucoup moins le modèle appris alors qu'elles sont censées être plus fiables que celles générées à partir des règles. Pour compenser cette sous-représentation, nous avons testé deux méthodes permettant d'augmenter leur impact sur le processus d'apprentissage.

La première méthode consiste simplement à suréchantillonner les données manuelles, en les reproduisant plusieurs fois à l'identique dans le corpus d'entrainement. Dans la seconde méthode, le corpus d'apprentissage reste le même (il contient donc les données issues des règles et celles annotées à la main en un seul exemplaire), mais nous contraignons l'algorithme d'apprentissage à considérer les données manuelles comme ensemble de validation. L'impact de ces deux méthodes est présentée dans la section suivante.

5 Évaluations

Pour les expérimentations, nous avons utilisé la bibliothèque RankLib[1], qui implémente plusieurs algorithmes d'apprentissage d'ordonnancement. Chaque algorithme cité dans la suite a été utilisé avec les paramètres définis par défaut dans la bibliothèque. D'une façon générale, nos efforts ne se sont pas concentrés sur l'optimisation des scores, notre objectif étant avant tout de montrer que le modèle s'améliore en prenant en compte les retours utilisateurs.

La table 1 montre les résultats obtenus pour l'ensemble des algorithmes testés et pour différentes configurations des données d'entrainement et de validation. De par la métrique d'évaluation choisie (ERR@3), les valeurs données dans cette section sont relativement basses en comparaison des métriques habituelles de la RI. En effet, comme expliqué plus haut, ce score représente la probabilité que l'utilisateur ait trouvé ce qu'il cherche dans les trois premières suggestions ; il n'est donc pas lié uniquement à la qualité de l'ordonnancement, mais aussi à la capacité du corpus à répondre au besoin en information. Pour mettre en évidence cette propriété de la métrique, nous avons calculé sa valeur maximale, c'est-à-dire la note qu'obtiendrait un oracle produisant un ordonnancement parfait.

Nous avons également construit une heuristique simple pour obtenir une *baseline*. Celle-ci renvoie toujours les trois mêmes suggestions, les plus populaires dans le corpus d'entrainement (la popularité est mesurée par le somme des notes attribuées). De par la très forte présence d'un nombre restreint de questions, cette *baseline* donne d'assez bons résultats, ce qui nous a incité à intégrer la notion de popularité des questions cibles dans les traits décrits plus haut.

Aucune tendance ne se dégage clairement de l'ajout des questions de la base de connaissances au corpus d'entrainement (deuxième colonne). Les résultats de l'apport des retours utilisateurs sont plus encourageants. Même si lorsque l'on ajoute simplement les données annotées à la main aux données d'entrainement, le score n'est amélioré que pour la moitié des algorithmes (troisième colonne), on constate une amélioration significative pour les deux algorithmes présentant les meilleures performances (Coordinate Ascent et

[1] https://sourceforge.net/p/lemur/wiki/RankLib/

LambdaMART, lignes en gras dans le tableau). De plus, la progression est plus marquée encore lorsque l'on donne artificiellement plus de poids à ces données. La quatrième colonne montre les résultats obtenus lorsque l'on fixe les données de validation comme étant les données annotées à la main. Enfin, les cinquième et sixième colonnes montrent les scores obtenus en dupliquant les données annotées à la main dans le corpus d'entrainement.

TABLE 1 – *Résultats obtenus pour l'ensemble des algorithmes testés et pour différentes configurations des données d'entrainement et de validation. Les valeurs montrent le score ERR@3 de chaque approche ; les scores en vert (resp. rouge) indiquent une amélioration (resp. dégradation)des résultats obtenus en augmentant le corpus d'apprentissage avec les retours utilisateurs.*

		apport des questions de la KB	apport des feedbacks utilisateurs			
				avec suréchantillonage		
données d'entrainement	règles	règles + questions kb	règles + questions kb + feedback		règles + questions kb + 3x feedbacks	règles + questions kb + 10x feedbacks
jeu de validation	20% du jeu d'entr.	20% du jeu d'entr.	20% du jeu d'entr.	feedback	20% du jeu d'entr.	20% du jeu d'entr.
MART	0,4319	0,4299	0,4299	0,4230	0,4293	0,422
RankNet	0,2623	0,2686	0,2794	0,2694	0,2626	0,2668
RankBoost	0,4147	0,4197	0,3908	0,3908	0,4179	0,3817
AdaRank	0,4495	0,4318	0,4203	0,4661	0,4407	0,4203
Coordinate Ascent	**0,4531**	**0,4547**	**0,4549**	**0,4549**	**0,4539**	**0,4547**
LambdaMART	**0,4548**	**0,4527**	**0,4575**	**0,4550**	**0,4845**	**0,4773**
ListNet	0,2603	0,2599	0,2597	0,2552	0,2597	0,2588
Random Forests	0,4216	0,4224	0,4218	0,4210	0,4227	0,4225
Popularité **(baseline)**	0,4030	0,3986	0,4023		0,4266	0,4514
Ordonnancement parfait **(upperline)**	0,6775					

6 Conclusion et perspectives

Nos expérimentations ont mis en évidence une augmentation de la qualité du modèle lorsque celui-ci prend en compte les retours utilisateurs. Ces résultats encourageants nous poussent à poursuivre nos travaux dans cette direction. Le système obtenu étant suffisamment mature, une perspective évidente est de le pousser en production, d'y intégrer progressivement les feedbacks utilisateurs et d'observer son évolution.

Un système en production doit faire face à une difficulté supplémentaire, introduite plus haut : l'utilisateur est susceptible, soit parce qu'il n'a pas compris, soit parce qu'il joue, de rentrer des phrases qui n'expriment aucun besoin en information. Il est alors essentiel, pour assurer la crédibilité du système, d'identifier correctement ces entrées afin de ne pas afficher la réponse à une question sans rapport. L'approche la plus simple est certainement d'entrainer un modèle de classification. Nous devrons probablement produire de nouveaux traits, plus pertinents pour cette tâche.

La substitution d'un composant par un autre entraîne toujours un risque de régression. L'identification et la mesure de ces régressions est un aspect critique lorsque l'on gère un système en production, d'autant plus lorsque, comme dans le scénario considéré, on veut remplacer un système de règles écrit à la main par un modèle issu d'un algorithme d'apprentissage automatique, difficile à ajuster une fois construit et souvent vu comme une boîte noire.

Les approches d'apprentissage actif telles que celle que nous présentons ici se marient en général très bien avec des algorithmes d'apprentissage en ligne (*online learning*), ceux-ci permettant d'intégrer progressivement de nouvelles données pour faire évoluer un modèle en temps réel. Dans les évaluations décrites plus haut, chaque modèle est à chaque fois construit intégralement à partir de l'ensemble des données d'apprentissage (en *batch*). Cela n'a posé aucun problème, étant donné la taille relativement réduite de notre corpus, mais nous voulons anticiper le passage à l'échelle en intégrant un algorithme d'apprentissage en ligne. Il faudra alors également penser à gérer le problème des traits représentant la popularité d'une question candidate, ceux-ci étant calculés sur l'ensemble du corpus.

Enfin, une intégration dans un champ de recherche avec une autocomplétion suggérant des questions proches veut épargner à l'utilisateur la nécessité de renseigner l'intégralité de sa question. Dans ce cas, la fonction de suggestion de questions doit être opérationnelle (même si de moins bonne qualité) dès les premiers caractères entrés par l'utilisateur. Cet aspect est déjà pris en compte par le composant d'analyse de questions, mais le réordonnancement des questions candidates, pour être pertinent, devra probablement être appris sur un corpus contenant des débuts de questions, que l'on pourrait construire en prenant les questions du corpus existant et en les tronquant à différentes tailles.

Références

ATTARDI, G., CISTERNINO, A., FORMICA, F., SIMI, M., TOMMASI, A., & ZAVATTARI, C. (2001). PiQASso: Pisa Question Answering System. In TREC.

BUNESCU, R., & HUANG, Y. (2010). LEARNING THE RELATIVE USEFULNESS OF QUESTIONS IN COMMUNITY QA. Conference on Empirical Methods in Natural Language Processing (pp. 97-107).

BURGES, C., SHAKED, T., RENSHAW, E., LAZIER, A., DEEDS, M., HAMILTON, N., & HULLENDER, G. (2005, August). Learning to rank using gradient descent. In Proceedings of the 22nd international conference on Machine learning (pp. 89-96). ACM.

CAO, Z., QIN, T., LIU, T. Y., TSAI, M. F., & LI, H. (2007, June). Learning to rank: from pairwise approach to listwise approach. In Proceedings of the 24th international conference on Machine learning (pp. 129-136). ACM.

CAO, Y., XU, J., LIU, T. Y., LI, H., HUANG, Y., & HON, H. W. (2006, August). Adapting ranking SVM to document retrieval. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 186-193). ACM.

CHAKRABARTI, S. (2004). Breaking through the syntax barrier: Searching with entities and relations. In Knowledge Discovery in Databases: PKDD 2004 (pp. 9-16). Springer Berlin Heidelberg.

CHAPELLE, O., METLZER, D., ZHANG, Y., & GRINSPAN, P. (2009, November). Expected reciprocal rank for graded relevance. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 621-630). ACM.

COSSOCK, D., & ZHANG, T. (2006, June). Subset ranking using regression. In Proceedings of the 19th annual conference on Learning Theory (pp. 605-619). Springer-Verlag.

CRAMMER, K., & SINGER, Y. PRANKING WITH RANKING. (2001). Advances in Neural Information Processing Systems, 14.

DAMLJANOVIĆ, D., AGATONOVIĆ, M., CUNNINGHAM, H., & BONTCHEVA, K. (2013). Improving habitability of natural language interfaces for querying ontologies with feedback and clarification dialogues. Web Semantics: Science, Services and Agents on the World Wide Web, 19, 1-21.

DEKLEVA, S. M. (1994). Is natural language querying practical?. ACM SIGMIS Database, 25(2).

DESERT, S. E. (1993). WESTLAW Is Natural v. Boolean searching: A performance study. J., 85, 713.

DITTENBACH, M., MERKL, D., & BERGER, H. (2003). A natural language query interface for tourism information. ENTER 2003: 10th International Conference on Information Technologies in Tourism.

FREUND, Y., IYER, R., SCHAPIRE, R. E., & SINGER, Y. (2003). An efficient boosting algorithm for combining preferences. The Journal of machine learning research, 4, 933-969.

HERBRICH, R., GRAEPEL, T., & OBERMAYER, K. (1999). Large margin rank boundaries for ordinal regression. Advances in neural information processing systems, 115-132.

HIRSCHMAN, L., & GAIZAUSKAS, R. (2001). Natural language question answering: the view from here. natural language engineering, 7(4), 275-300.

HOVY, E. H., GERBER, L., HERMJAKOB, U., JUNK, M., & LIN, C. Y. (2000, November). Question Answering in Webclopedia. In TREC (Vol. 52, pp. 53-56).

JÄRVELIN, K., & KEKÄLÄINEN, J. (2002). Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS), 20(4), 422-446.

KAUFMANN, E., & BERNSTEIN, A. (2010). Evaluating the usability of natural language query languages and interfaces to Semantic Web knowledge bases. Web Semantics: Science, Services and Agents on the World Wide Web, 8(4), 377-393.

LAURENT, D., SEGUELA, P., & NEGRE, S. (2005). QRISTAL, système de Questions-Réponses. TALN & RECITAL, 1, 53-62.

MOLDOVAN, D. I., HARABAGIU, S. M., GIRJU, R., MORARESCU, P., LACATUSU, V. F., NOVISCHI, A., & BOLOHAN, O. (2002, November). LCC Tools for Question Answering. In TREC.

MOLDOVAN, D., HARABAGIU, S., PASCA, M., MIHALCEA, R., GOODRUM, R., GIRJU, R., & LASSO, V. R. (1999, November). A tool for surfing the answer net. In TREC-8 Proceedings.

LI, H. (2011). A short introduction to learning to rank. IEICE TRANSACTIONS on Information and Systems, 94(10), 1854-1862.

LI, H. (2014). Learning to rank for information retrieval and natural language processing. Synthesis Lectures on Human Language Technologies, 7(3), 1-121.

LI, P., WU, Q., & BURGES, C. J. (2007). Mcrank: Learning to rank using multiple classification and gradient boosting. In Advances in neural information processing systems (pp. 897-904).

LINCKELS, S., & MEINEL, C. (2005). A simple solution for an intelligent librarian system. In IADIS AC

LITKOWSKI, K. C. (2000, November). Syntactic Clues and Lexical Resources in Question-Answering. In TREC.

LIU, T. Y. (2009). Learning to rank for information retrieval. Foundations and Trends in Information Retrieval, 3(3), 225-331.

MIN, W., XIAOYU, Z., & MICHELLE, D. (2003). Question Answering By Pattern Matching, Web-Proofing, Semantic Form Proofing. In Proceedings of the TREC-12 Conference. Maryland: NIST Special Publication (pp. 165-169).

REICHERT, M., LINCKELS, S., MEINEL, C., & ENGEL, T. (2005, December). Student's Perception of a Semantic Search Engine. In CELDA (pp. 139-147).

SHASHUA, A., & LEVIN, A. (2002). Ranking with large margin principle: Two approaches. In Advances in neural information processing systems (pp. 937-944).

SRIHARI, R., & LI, W. (1999). Information extraction supported question answering. CYMFONY NET INC WILLIAMSVILLE NY.

SURDEANU, M., CIARAMITA, M., & ZARAGOZA, H. (2008). Learning to Rank Answers on Large Online QA Collections. In ACL (Vol. 8, pp. 719-727).

TAYLOR, M., GUIVER, J., ROBERTSON, S., & MINKA, T. (2008, February). Softrank: optimizing non-smooth rank metrics. In Proceedings of the 2008 International Conference on Web Search and Data Mining (pp. 77-86). ACM.

THOMPSON, C. W., PAZANDAK, P., & TENNANT, H. R. (2005). Talk to your semantic web. IEEE Internet Computing, 9(6), 75.

VOORHEES, E. M., & TICE, D. M. (1999, November). The TREC-8 Question Answering Track Evaluation. In TREC (Vol. 1999, p. 82).

WU, Q., BURGES, C. J., SVORE, K. M., & GAO, J. (2010). Adapting boosting for information retrieval measures. Information Retrieval, 13(3), 254-270.

XIA, F., LIU, T. Y., WANG, J., ZHANG, W., & LI, H. (2008, July). Listwise approach to learning to rank: theory and algorithm. In Proceedings of the 25th international conference on Machine learning (pp. 1192-1199). ACM.

XU, J., & LI, H. (2007, July). Adarank: a boosting algorithm for information retrieval. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 391-398). ACM.

YUE, Y., FINLEY, T., RADLINSKI, F., & JOACHIMS, T. (2007, July). A support vector method for optimizing average precision. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 271-278). ACM.

ZHENG, Z., ZHA, H., ZHANG, T., CHAPELLE, O., CHEN, K., & SUN, G. (2008). A general boosting method and its application to learning ranking functions for web search. In Advances in neural information processing systems (pp. 1697-1704).

Modèle unifié pour la recherche d'information sémantique

Ines Bannour, Haïfa Zargayouna, Adeline Nazarenko

LABORATOIRE D'INFORMATIQUE DE PARIS NORD (LIPN, UMR 7030)
Université Paris 13 – Sorbonne Paris Cité & CNRS
Email: prenom.nom@lipn.univ-paris13.fr

Résumé : Un modèle documentaire permet de définir les unités d'indexation (mots, termes, etc.) et de les relier aux documents dans lesquels elles apparaissent. Il permet également de définir les liens entre documents ou portions de documents (ex. citation). Les modèles documentaires sont généralement exploités en recherche d'information pour la représentation des documents et des requêtes et ils autorisent des calculs de pertinence numériques fondés sur la répartition des unités d'indexation dans la collection de documents.

Le modèle sémantique définit les unités sémantiques (concepts, instances de concepts, etc.) qui peuvent être reliées par des relations (relations hiérarchiques ou rôles). En recherche d'information, ils permettent d'aller au-delà des mots et de raisonner au niveau des concepts ou instances de concepts.

Nous proposons d'unifier les modèles documentaire et sémantique dans un unique réseau sémantico-documentaire pour représenter l'ensemble des propriétés, qu'elles soient numériques ou symboliques. La propagation d'activation est utilisée pour propager l'information de pertinence de proche en proche sur le graphe, depuis les éléments de la requête utilisateur.

Dans cet article, nous présentons notre modèle et les requêtes qu'il permet de prendre en compte. Nous présentons également des résultats sur des expérimentations préliminaires effectuées sur un banc de test de l'état de l'art. Nous obtenons des performances comparables à l'état de l'art pour des modèles simples, ce qui fait espérer des marges de progrès avec l'introduction de la sémantique.

Mots-clés : Documents, ontologie, annotation sémantique, graphe, propagation d'activation, recherche d'information

1 Introduction

L'avènement du Web, des moteurs de recherche et du Web Sémantique (WS) a décuplé l'information disponible et les moyens d'accéder à cette information. Les modèles sous-jacents sont cependant hétérogènes : les moteurs de recherche reposent essentiellement sur la fréquence des mots et l'analyse de leurs distributions dans les documents ; la recherche d'information sémantique (RIS) exploite à l'inverse des connaissances sémantiques généralement consignées dans des ressources comme les ontologies ou les thesaurus Zargayouna *et al.* (2015).

Des travaux récents (Castells *et al.*, 2007; Bhagdev *et al.*, 2008; Fernández *et al.*, 2011) proposent de combiner différents espaces d'indexation qui pour exploiter au mieux les modèles sémantiques tout en gardant une représentation classique du modèle documentaire tel que le modèle vectoriel défini par Salton *et al.* (1975).

Le défi aujourd'hui consiste à proposer un modèle unifié qui permette à l'utilisateur d'avoir accès à l'ensemble de ces fonctionnalités dans un unique système d'accès à l'information. Il doit pouvoir interroger une base documentaire à l'aide de mots-clefs ou de concepts mais aussi retrouver des documents similaires à un texte source voire les concepts associés à un ensemble de termes.

Dans ce travail nous proposons d'exploiter à la fois les connaissances sémantiques des ontologies du WS et les caractéristiques distributionnelles largement éprouvées en RI. Nous intégrons pour cela les relations sémantiques des ontologies et les relations termes-documents de

la RI traditionnelle dans un unique modèle de graphe pondéré et nous modélisons la fonction de correspondance requête-résultats sous la forme d'un mécanisme de propagation d'activation dans le graphe.

La suite de cet article est organisée comme suit : la section 2 présente le modèle proposé, la section 3 présente les travaux en propagation d'activation pour la RI, la section 4 donne les premiers résultats obtenus.

2 Modèle

Un modèle de Recherche d'Information propose une manière unifiée de représenter les requêtes et les documents ainsi qu'une *fonction de correspondance* qui associe des scores aux couples requête-document permettant ainsi de trier les documents en fonction de la requête.

Notre approche repose sur un modèle de graphe qui permet de représenter dans un modèle unique la base documentaire, avec notamment les relations termes-documents, et le réseau sémantique qui comporte par exemple une structure de concepts et des associations entre termes et concepts. La correspondance entre les requêtes et les documents du graphe est calculée par un mécanisme de propagation d'activation sur ce graphe.

2.1 Réseau sémantico-documentaire

Nous proposons de représenter le modèle documentaire et le modèle sémantique qui lui est associé sous la forme d'un unique réseau sémantico-documentaire. Cette structure permet d'introduire différents types de noeuds et différents types de relations selon ce qu'on souhaite représenter.

Nous proposons de prendre en compte trois types de noeuds : les *noeuds documents* représentent tous les documents de la collection documentaire ; les *noeuds termes* représentent le vocabulaire de la collection documentaire et les *noeuds concepts* représentent les concepts et instances de l'ontologie associée à la collection documentaire.

Ces noeuds sont reliés par 5 types de relations qui peuvent porter des propriétés :

— les *relations d'occurrence* sont des relations entre termes et documents qui traduisent le fait qu'un terme apparaît dans un document ; une propriété de fréquence peut naturellement être associée à ces relations ;

— les *relations d'intertextualité* sont des relations entre documents, comme par exemple les relations de citation ; ces liens peuvent être typés, les relations de citation n'étant pas les seules à être intéressantes à prendre en compte [1] ;

— les *relations terminologiques* sont des relations entre termes et concepts qui indiquent quels termes sont les labels de quels concepts : dans les ressources sémantiques dotées d'une composante terminologique, le fait qu'un terme soit relié à plusieurs concepts traduit son ambiguïté ; un concept peut également avoir plusieurs labels qui le dénotent, certains pouvant être des termes « préférés » ;

— les *relations d'annotation* sont des relations entre documents et concepts associant des concepts ou des catégories comme méta-données à des documents et qui sont souvent

1. Dans le domaine juridique, Mimouni *et al.* (2014) évoquent par exemple la relation de transposition entre une directive européenne et un texte réglementaire ou législatif national.

issues d'un travail d'annotation sémantique des documents ;
— les *relations ontologiques* sont des relations entre concepts (ou concepts et instances) qui peuvent représenter aussi bien les rôles que les liens hiérarchiques.

2.2 Graphe pondéré

Ce réseau sémantico-documentaire peut être représenté sous la forme d'un graphe pondéré. Ce graphe $G = \langle N, R \subseteq N \times \mathbb{R} \times N \rangle$ est constitué d'un ensemble de noeuds ($N = N_d \uplus N_t \uplus N_c$) et d'arcs qui sont orientés et pondérés ($R = R_{occ} \uplus R_{int} \uplus R_{ter} \uplus R_{ann} \uplus R_{ont}$).

Chacune de ces relations peut porter un poids qui reflète son importance dans le graphe : le poids d'une relation d'occurrence peut représenter la fréquence d'un terme dans un document ; le poids d'une relation terminologique peut permettre de distinguer le label « préféré » d'un concept par rapport aux autres termes qui lui sont associés ; etc. Nous n'entrons pas ici dans le détail du calcul de ces poids, considérant que différents paramétrages sont possibles, depuis un graphe booléen (sans poids) à un graphe entièrement pondéré, qu'ils reflètent différents choix de modélisation mais qu'ils sont tous compatibles avec le modèle à base de graphe que nous proposons.

Le graphe pondéré peut être interrogé de plusieurs manières selon que la requête comporte des termes (comme en RI traditionnelle), des concepts (comme en RIS), des documents (par ex. pour une recherche à base d'exemples) ou une combinaison de ces différents types d'éléments : $Q = \{t_1, t_2, \ldots, C_1, C_2, \ldots, D_1, D_2, \ldots\}$. Les réponses attendues peuvent également être de différents types (documents, termes, concepts). Le modèle unifié à base de graphe permet ainsi de prendre en compte diverses formes de requêtes et de proposer différents types de résultats, sans avoir à changer de système d'accès à l'information ou de langage d'interrogation.

2.3 Propagation d'activation

La propagation d'activation est un processus qui permet de propager une information de proche en proche sur un graphe. Ce mécanisme repose sur des valeurs d'activation associées aux noeuds du graphe : au départ les noeuds qui correspondent à la requête ont des valeurs d'activation positives et les autres noeuds sont neutres ; le processus de propagation est ainsi déclenché ; quand celui-ci s'arrête, les valeurs d'activation obtenues sur les noeuds du graphe [2] déterminent l'ordre de pertinence des noeuds de ce graphe au regard de la requête initiale.

La propagation à proprement parler consiste à 1) sélectionner les noeuds à activer parmi les noeuds dont la valeur d'activité est non nulle et qui n'ont pas encore été activés, puis à 2) propager l'activité de ces noeuds à leurs voisins et les désactiver, ce processus étant itéré jusqu'à ce que plus aucun noeud ne puisse être sélectionné.

On comprend que la propagation s'applique aux valeurs d'activation qui sont mises à jour par « contagion » sur les voisins mais qu'elle est aussi contrôlée par l'état des noeuds, lequel évolue au cours du processus, un noeud étant tour à tour *inactif*, *activé* et *désactivé*.

La valeur d'activation du noeud i à la k^{ime} itération est calculée de la manière suivante :

$$a_k(i) = a_{k-1}(i) + \sum_{j \in pred(i) \cup actif(k-1)} a_{k-1}(j) * w(j,i) * 1/deg(j)$$

2. Les noeuds peuvent être filtrés si on veut restreindre le type de résultat.

Elle dépend de la structure du graphe, à savoir les prédécesseurs du noeud i ($j \in pred(i)$) et le degré de ces noeuds (deg(j)) mais aussi de l'état des noeuds du graphe, seuls les actifs à l'itération $(k-1)$ étant pris en compte ($j \in actif(k-1)$). On note que la fonction d'activation pour un noeud est croissante : un noeud désactivé ne peut plus être réactivé mais sa valeur d'activation peut continuer à croître sous l'effet de ses voisins. La propagation d'activation s'arrête quand il n'y a plus de noeuds actifs : il ne reste que des noeuds déjà désactivés ou des noeuds inactifs qui ne peuvent être atteint par la propagation d'activation.

3 État de l'art

Les approches à base de graphes ont pris beaucoup d'importance en recherche d'information où les méthodes des marches aléatoires sont répandues depuis PageRank (Brin & Page, 1998). Même si le modèle mathématique de la propagation d'activation est moins solidement fondé que celui des algorithmes à base de marches aléatoires, il a une complexité moindre et s'adapte mieux au cadre d'une fonction de correspondance, qui est dirigée par la requête.

L'application de la propagation d'activation en recherche d'information n'est pas récente (Preece, 1981; Cohen & Kjeldsen, 1987; Croft *et al.*, 1988; Salton & Buckley, 1988; Savoy, 1992). Crestani (1997) présente un état de l'art sur les travaux en recherche d'information qui ont proposé l'utilisation de la propagation d'activation dans des réseaux associatifs ou des réseaux sémantiques.

Plusieurs travaux en WS proposent de peupler la base de connaissances avec des documents mais la représentation des documents dans une base de connaissances ne permet pas de mettre en place une recherche documentaire. Les bases de connaissances sont utiles pour une recherche précise avec une vision orientée données qui nécessite de connaître la structure de la base. Les graphes de connaissances, tels que les graphes RDF ou les graphes conceptuels ne sont pas adaptés à la RI car ils modélisent essentiellement des données symboliques. En effet, il est difficile de mettre en place des calculs distributionnels car ils ne permettent pas de prendre en compte des informations telles que les fréquences et le nombre d'occurrences.

Le modèle de propagation d'activation que nous proposons permet de reproduire la RI classique et d'exploiter des informations ontologiques difficiles à mettre en place dans une base de connaissances.

4 Expérimentations

Nous avons enrichi la plateforme Terrier RIS proposée par Bannour & Zargayouna (2012) par la plateforme d'analyse de graphes JUNG (*Java Universal Network/Graph Framework*) [3]. La plateforme JUNG permet de modéliser différents types de graphes (orienté, non orienté, etc.). Elle implémente également de nombreux algorithmes sur les graphes.

Les premières expérimentations ont porté sur le corpus de recettes de cuisine exploité par Bannour & Zargayouna (2012). Le corpus est composé de 1 489 recettes de cuisines et de 4 requêtes avec leurs jugements de pertinence [4].

3. http ://jung.sourceforge.net/

4. Une mise à jour des jugements de pertinence a été effectuée. Exemple, la requête 1 : Cook an Asian soup with leek (6 documents pertinents).

La *baseline* consiste en une implémentation classique du modèle vectoriel (Salton *et al.*, 1975) avec la formule de pondération TF-IDF (*Term Frequency-Inverse Document Frequency*).

Nous avons voulu, dans un premier temps, nous assurer que notre modèle permet de simuler un comportement classique de RI dans le cas où le modèle représente uniquement les termes, les documents et les liens entre eux. Le poids des liens terme-document et document-terme est calculé par :

$$w(t, doc) = \frac{tf(t, d)}{maxTF(doc)}$$

tel que $maxTF(doc)$ présente la fréquence maximale d'un terme dans le document d.

Nous présentons les résultats en termes de MAP et R-PREC : la MAP (*Mean Average Precision*) est la moyenne de la précision obtenue après chaque document pertinent retourné ; la R-PREC (*R-Precision*) est la précision après R documents pertinents retournés, R étant le nombre de documents pertinents. Les résultats de notre approche (GraphTerm) sont meilleurs que ceux de la base line (TF.IDF) : le tableau 4 fait apparaître une augmentation de près de 10% et, pour la requête 1 par exemple, une nette amélioration en termes de MAP et R-précision.

Méthode	MAP				R-PREC			
	R1	R2	R3	R4	R1	R2	R3	R4
TF.IDF	0.25	0.6	0.15	0.47	0.33	0.66	0.0	0.5
Total	0.36				0.37			
GraphTerm	0.5	0.78	0.12	0.5	0.66	0.66	0.0	0.5
Total	0.47				0.46			

FIGURE 1 – Tableau récapitulatif des résultats

Ces résultats, même s'ils sont préliminaires, montre que notre modèle donne des résultats similaires à ceux qu'on peut obtenir avec un modèle éprouvé, en recherche d'information classique. Brouard (2013) a présenté des résultats équivalents mais avec un modèle qui exploite d'une manière différenciée la couche documents de la couche termes en proposant deux formules de propagation différentes.

5 Conclusion

Nous avons proposé d'unifier les modèles documentaire et sémantique dans un unique réseau sémantico-documentaire qui permet de représenter l'ensemble des propriétés du modèle documentaire et du modèle sémantique. Ce réseau est représenté sous forme de graphe pondéré. L'intérêt de cette représentation est de combiner des propriétés numériques et symboliques. Nous proposons d'appliquer une propagation d'activation qui permet de propager l'information de pertinence de proche en proche sur le graphe. Les premières expérimentations montrent que nous obtenons de bonnes performances par rapport à l'état de l'art pour des modèles simples, ce qui fait espérer des marges de progrès avec l'introduction de la sémantique. Des expérimentations à plus grande échelle sont en cours. Ces expérimentations vont permettre de calibrer les formules de propagation, de mettre en place les heuristiques de recherche nécessaires, et d'étudier l'impact des structures de connaissances sur la propagation.

Remerciements

Ce travail a bénéficié partiellement d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'Avenir » portant la référence ANR-10-LABX-0083.

Références

BANNOUR I. & ZARGAYOUNA H. (2012). Une plate-forme open-source de recherche d'information sémantique. In *COnférence en Recherche d'Information et Applications (CORIA)*, p. 167–178.

BHAGDEV R., CHAPMAN S., CIRAVEGNA F., LANFRANCHI V. & PETRELLI D. (2008). Hybrid search : Effectively combining keywords and semantic searches. In *Proceedings of the 5th European Semantic Web Conference, ESWC*, p. 554–568.

BRIN S. & PAGE L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web (WWW7)*, p. 107–117.

BROUARD C. (2013). Comparaison du modèle vectoriel et de la pondération tf*idf associée avec une méthode de propagation d'activation. In *CORIA*, p. 217–226.

CASTELLS P., FERNANDEZ M. & VALLET D. (2007). An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Transactions on Know. and Data Eng.*, **19**(2), 261–272.

COHEN P. R. & KJELDSEN R. (1987). Information retrieval by constrained spreading activation in semantic networks. *Inf. Process. Manage.*, **23**(4), 255–268.

CRESTANI F. (1997). Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, **11**(6), 453–482.

CROFT W. B., LUCIA T. J. & COHEN P. R. (1988). Retrieving documents by plausible inference : A priliminary study. In *Proceedings of the 11th Annual International ACM SIGIR*, SIGIR '88, p. 481–494.

FERNÁNDEZ M., CANTADOR I., LÓPEZ V., VALLET D., CASTELLS P. & MOTTA E. (2011). Semantically enhanced information retrieval : an ontology-based approach. *Web Semantics : Science, Services and Agents on the World Wide Web*, **9**(4), 434–452.

MIMOUNI N., NAZARENKO A., PAUL È. & SALOTTI S. (2014). Towards graph-based and semantic search in legal information access systems. In *Legal Knowledge and Information Systems - JURIX*, volume 271, p. 163–168.

PREECE S. (1981). *A Spreading Activation Network Model for Information Retrieval*. University of Illinois at Urbana-Champaign.

SALTON G. & BUCKLEY C. (1988). On the use of spreading activation methods in automatic information. In *Proceedings of the 11th Annual International ACM SIGIR*, SIGIR '88, p. 147–160.

SALTON G., WONG A. & YANG C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, **18**(11), 613–620.

SAVOY J. (1992). Bayesian inference networks and spreading activation in hypertext systems. *Information Processing Management*, **28**(3), 389 – 406.

ZARGAYOUNA H., ROUSSEY C. & CHEVALLET J. P. (2015). Recherche d'information sémantique : état des lieux. *TAL (Traitement Automatique des Langues)*, **56**(3), 49–73.

Recherche collaborative de documents : comparaison assistance humaine/automatique

Jean-Baptiste Louvet[1], Guillaume Dubuisson Duplessis[2], Nathalie Chaignaud[1], Jean-Philippe Kotowicz[1,3], Laurent Vercouter[1]

[1] Normandie Univ, INSA Rouen, LITIS, 76000 Rouen, France
{jeanbaptiste.louvet, nathalie.chaignaud, laurent.vercouter}@insa-rouen.fr

[2] LIMSI-CNRS, Paris
gdubuisson@limsi.fr

[3] UEMF, INSA Euro-Méditerranée, Fès, Maroc
j.kotowicz@insa.ueuromed.org

Résumé : Partant du postulat que l'étude de l'interaction humain-humain peut servir à modéliser l'interaction humain-machine, nous étudions un corpus de dialogues de recherche collaborative de documents. De cette analyse est né un modèle de la tâche dans le cadre d'une interaction humain-humain. L'utilité de ce modèle est discuté pour le cas d'une interaction humain-machine.

Mots-clés : Recherche de documents, interaction humain-humain, interaction humain-machine

1 Introduction

Nous nous intéressons au système d'indexation de connaissances médicales CISMeF (Darmoni *et al.*, 2000). Partant de l'étude d'un corpus de dialogues d'assistance à la recherche de documents (RD), nous définissons un modèle de la tâche de recherche collaborative de documents dans le cadre d'une interaction humain-humain. Nous montrons en quoi ce modèle de collaboration humain-humain (h-h) peut servir de base de discussion pour le cas d'une interaction humain-machine (h-m).

2 Recherche d'information

Cette section présente quelques modèles du processus de recherche d'information (RI) réalisé par un individu isolé pouvant s'appliquer à la RD. Un état de l'art plus complet est disponible dans les travaux de Dubuisson Duplessis (Dubuisson Duplessis, 2014).

On conçoit généralement la RI comme un processus de résolution de problème (Marchionini, 1989) impliquant un *chercheur* ayant un *besoin d'information identifié*. Le problème est alors de combler ce manque d'information. Le besoin d'information spécifié, le chercheur choisit un *plan de recherche* qu'il exécute au travers de la recherche en elle-même. Il évalue les résultats trouvés pour éventuellement réitérer le processus entier. La RI est un *processus itératif* (Sutcliffe & Ennis, 1998; Broder, 2002; Marchionini & White, 2007).

Le modèle standard est limité sur deux aspects. D'un côté, le *besoin d'information* de ce processus est vu comme *statique*. De l'autre, le chercheur est vu comme raffinant successivement sa requête jusqu'à tomber sur un ensemble de documents répondant à son besoin d'information *initial*. Certaines études ont démontré au contraire que le besoin d'information n'est pas statique

et que l'objectif n'est pas de déterminer une requête unique retournant un ensemble de documents répondant au besoin d'information (Bates, 1989; O'Day & Jeffries, 1993).

3 Analyse du processus de recherche collaborative h-h de documents à partir du corpus

De manière à comprendre l'aspect collaboratif d'un processus de RD d'un utilisateur assisté par un expert et à analyser les stratégies mises en place par ce dernier, nous avons mené une étude sur une interaction h-h de RD médicale. Cette étude est fondée sur l'analyse du corpus h-h recueilli lors du projet COGNI-CISMEF (Loisel, 2008; Chaignaud *et al.*, 2010).

3.1 Présentation du corpus COGNI-CISMEF

Le corpus est constitué de dialogues d'assistance sur une tâche de RD entre un expert et un utilisateur : l'expert se retrouve en situation de co-présence avec l'utilisateur qui apporte une formulation de son besoin d'information. L'expert dispose d'un accès au portail CISMEF et est chargé de mener la recherche en coopération avec l'utilisateur. Le portail CISMEF propose une interface graphique et un langage de requête permettant de décomposer une requête en éléments issus d'un lexique contrôlé. La terminologie CISMEF comporte des mots-clés, des qualificatifs , des méta-termes et des types de ressources. Le système s'est étoffé en offrant la possibilité d'effectuer des requêtes étendues mais l'utilisateur est souvent peu enclin à l'utilisation de requêtes complexes pourtant plus efficaces.

L'expérimentation a été menée auprès de membres volontaires (enseignants-chercheurs, étudiants, administratifs...) qui soumettaient une requête à un expert. Deux chercheurs du projet (un informaticien et une psycholinguiste), s'étant formés à la terminologie CISMEF, jouaient le rôle d'expert. Le corpus est constitué des retranscriptions des 21 dialogues issus de cette expérimentation et contient environ 37 000 mots.

3.2 Phases de la recherche de documents

L'analyse du corpus a permis d'identifier et de caractériser les différentes phases des dialogues du corpus COGNI-CISMEF (Dubuisson Duplessis, 2014), qui jouent chacune un rôle dans l'avancement de la tâche. Cinq phases ont été distinguées :

— la verbalisation : c'est l'établissement du sujet de la recherche entre l'utilisateur et l'expert. Elle commence par une formulation de la demande de l'utilisateur et peut être suivie de précisions spontanées de la part de celui-ci. L'expert peut alors commencer la construction de la requête s'il considère que la formulation est suffisante, demander des précisions si elle lui semble insuffisante ou tenter de reformuler la demande de l'utilisateur ;

— la construction de la requête : c'est la recherche de termes dans la terminologie CISMEF correspondant à la verbalisation de l'utilisateur. Il s'agit pour les interlocuteurs de trouver de façon collaborative un alignement entre la terminologie métier et le vocabulaire usuel de l'utilisateur pour ensuite remplir le formulaire de recherche ;

— le lancement de la requête : c'est simplement l'exécution de la requête courante par l'expert. Cette phase est souvent réalisée de manière implicite ;

— l'évaluation des résultats : l'expert évalue les résultats retournés par la requête. Si ceux-ci ne lui semblent pas satisfaisants, il décide de directement réparer la requête. Sinon, il les

présente à l'utilisateur. Si celui-ci les juge satisfaisants, le but est atteint et la recherche se termine ; s'il les juge partiellement satisfaisants ou insatisfaisants, il faut réparer la requête ; suite à une évaluation négative, il est aussi possible de terminer sur un abandon ;

— la réparation de la requête : l'expert et l'utilisateur tentent de mettre en place des tactiques pour modifier la requête tout en respectant le besoin d'information. Trois tactiques ont été observées : la précision, la reformulation et la généralisation. Cependant, ces tactiques ne sont pas mutuellement exclusives : il est possible de combiner une précision ou une généralisation avec une reformulation.

En plus de ces phases, une phase d'ouverture et de clôture ont été observés. La phase d'ouverture est facultative et consiste en de simples salutations, tandis que la phase de clôture fait apparaître des propositions d'une nouvelle recherche.

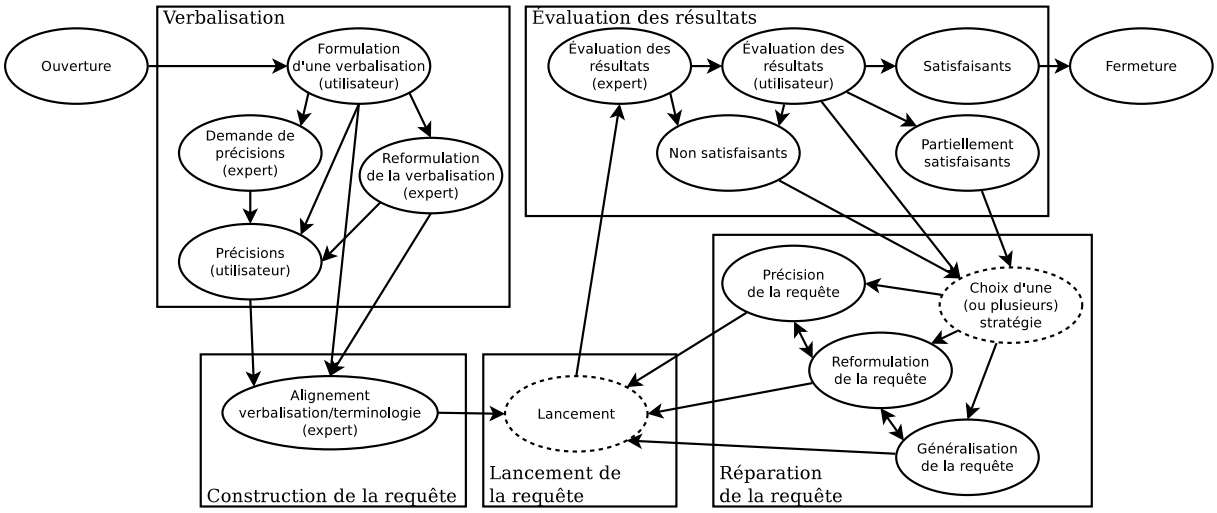


FIGURE 1 – Scénario issu de l'analyse du corpus représentant les enchaînements des phases de la tâche de RD.

3.3 Scénario issu de l'analyse du corpus

L'analyse de ce corpus a montré que cette tâche de RD est un processus itératif, opportuniste, stratégique et interactif (Dubuisson Duplessis, 2014; Bates, 1990). L'aspect itératif de ce processus est illustré par la répétition du motif lancement/évaluation/réparation de la requête.

Le scénario représenté par la Figure 1 synthétise les phases ainsi que les enchaînements possibles. Les états en pointillés sont les actions qui peuvent être réalisées de manière implicite par les participants à l'interaction. La boucle lancement/évaluation/réparation est bien présente.

4 Recherche collaborative de documents lors d'une interaction h-m

Nous nous intéressons maintenant aux différences entre une interaction h-h et une interaction h-m pour la RD collaborative. Nous comparons ici les capacités d'un expert humain et d'un système d'assistance dans une tâche d'assistance de formulation de requête.

4.1 Cadre de l'interaction h-m

Le cadre d'interaction h-m donne à l'utilisateur la possibilité de diriger la recherche et de modifier la requête sans l'accord du système. C'est un changement important car il donne l'ascendant à l'utilisateur sur l'interaction. Il implique une restriction des libertés du système par rapport à celles de l'expert humain. En effet, le système peut prendre des initiatives pour faire évoluer la requête en proposant des modifications qui ne seront effectives que si elles sont acceptées par l'utilisateur. Cette inversion des droits de modification de la requête permet cependant à chaque protagoniste de prendre part à l'interaction : l'initiative mixte est conservée.

Un autre changement est introduit : pour exposer son besoin d'information au système, l'utilisateur emploiera des mots libres ou éventuellement des phrases courtes avec utilisation d'outils TAL simples de la part du système (extraction de noms…).

4.2 Phase de verbalisation

Dans l'interaction h-h, c'est un moment d'interaction riche et complexe rendant explicite le besoin d'information de l'utilisateur. Cela nécessite de nombreuses compétences et connaissances de la part de l'expert : compréhension de l'expression verbale de l'utilisateur, connaissances générales du domaine d'application…

Dans le cadre h-m, cette phase sera forcément appauvrie. Un système d'assistance n'aura pas les capacités de TAL dont dispose l'expert humain et l'utilisateur devra donc entrer des mots libres (ou éventuellement des phrases très simples avec extraction de noms). Le système pourra toutefois demander des clarifications sur ces mots libres s'ils sont jugés trop peu nombreux.

L'expert humain est donc à son avantage dans la phase de verbalisation, celle-ci nécessitant des capacités cognitives difficiles à mettre en œuvre pour un système.

4.3 Phase de construction de la requête

Cette phase a pour objectif d'aligner la verbalisation de l'enquêté avec la terminologie CISMEF. Dans le meilleur des cas, un ou plusieurs termes correspondent directement avec la terminologie. Pour le système, cet alignement sera fait instantanément et le système pourra proposer à l'utilisateur d'ajouter à la requête les termes reconnus. Cette tâche est plus difficile pour l'expert humain car il doit extraire de la verbalisation les « termes importants » puis les saisir dans l'interface de CISMEF pour vérifier qu'ils correspondent ou non à un terme de la terminologie. Cela peut s'avérer fastidieux s'il y a plusieurs « termes importants ».

Dans le cas où les termes de la verbalisation ne correspondent pas à la terminologie, l'expert humain et l'utilisateur collaborent en parcourant celle-ci afin de trouver une correspondance. De même, le système peut demander à l'utilisateur de parcourir la terminologie pour proposer des termes. Il s'agit donc, pour l'humain ou le système, de faire appel au jugement de l'utilisateur.

Cependant, le système dispose des dictionnaires de synonymes, hyponymes, hyperonymes et de liens « voir aussi » de CISMEF permettant de trouver des termes proches de la verbalisation de l'utilisateur. De plus, il est capable de conserver les sessions de RD précédentes et peut donc aussi tirer parti de cette connaissance (liens entre besoins d'information, requêtes et documents) pour trouver des termes connexes à la recherche courante (El Guedria & Vercouter, 2015).

4.4 Phase de lancement de la requête

Le lancement de la requête arrive après la phase de formulation ou de réparation de la requête. Quand il est à l'initiative de l'expert, celui-ci considère que la requête est suffisamment complète pour apporter des résultats pertinents au besoin d'information de l'utilisateur. Cependant, il arrive que les résultats de la requête soient rejetés après une évaluation sommaire, dans le cas où il y a trop ou trop peu de documents ou quand ceux-ci sont manifestement hors sujet.

De son côté le système d'assistance a l'avantage de pouvoir lancer « en privé » (c-à-d sans que l'utilisateur ne le sache) les requêtes. L'intérêt est que, avant de proposer une modification à l'utilisateur, il pourra tester celle-ci pour déterminer si elle apporte des résultats intéressants ou non.

4.5 Phase d'évaluation des résultats

Dans une interaction h-h, cette phase commence par une évaluation des résultats par l'expert avant de les présenter à l'utilisateur pour que celui-ci juge leur pertinence. Cette étape est supprimée dans le cadre d'une interaction h-m car cette évaluation aura été faite en privé par le système. C'est un avantage pour le système, lui permettant d'anticiper sur l'intérêt de la requête. Cependant, le système est limité par sa capacité à donner une évaluation correcte des résultats. Là où l'expert humain possède des compétences cognitives pour rejeter ou non des résultats, le système se limitera à quelques paramètres descriptifs : nombre, score, métadonnées…

Dans le cas h-h et dans le cas h-m, le jugement est réalisé en dernier ressort par l'utilisateur.

4.6 Phase de réparation de la requête

La première étape dans une réparation de requête est l'identification de la tactique à mettre en place pour améliorer la requête. Le choix d'une tactique paraît assez simple : si la requête retourne trop de résultats, il faut la préciser ; si elle n'en retourne pas assez, il faut la généraliser et si elle en retourne un nombre suffisant mais que les documents ne sont pas satisfaisants pour l'utilisateur, il faut la reformuler. Cependant, ces tactiques peuvent être combinées.

Une précision nécessite un ajout de termes à la requête et les moyens employés sont proches de ceux utilisés lors de la phase de construction de la première requête. Une généralisation nécessite la suppression de termes de la requête. Grâce à ses capacités combinatoires, le système peut tester différentes possibilités de suppression et ainsi déterminer laquelle sera la plus avantageuse.

Pour la reformulation, le système peut faire appel aux ressources lexicales de l'application sur les termes de la requête (Audeh *et al.*, 2013; Soualmia, 2004) et il pourra faire appel à l'utilisateur pour proposer des reformulation de certains termes de la requête .

5 Conclusion et perspectives

Dans cet article, nous avons étudié une situation collaborative h-h de RD. Un modèle de la tâche entre un expert humain et un utilisateur est décrit sous la forme d'un scénario. À partir de ce modèle, nous avons comparé les capacités mises en jeu par l'expert humain et celles d'un système d'assistance à la formulation de requête.

Notre objectif est de concevoir un système collaboratif pour la RD capable d'interagir avec l'utilisateur et de lui proposer des stratégies coopératives. Le système doit pouvoir analyser

l'objectif de l'utilisateur et lui proposer des solutions pour faire évoluer l'état de sa recherche. Il doit présenter des exemples, des aides, des corrections ou des clarifications. Il doit accompagner l'interlocuteur jusqu'à trouver une solution en élargissant son but initial si nécessaire.

Une étude des aspects dialogiques de ce corpus a été proposée (Dubuisson Duplessis *et al.*, 2013, 2015). Elle montre que le modèle conventionnel des jeux de dialogue est bien adapté pour modéliser les niveaux interprétatif et génératif du dialogue. Nous travaillons actuellement sur les liens entre le modèle du dialogue et le modèle de la tâche.

Références

AUDEH B., BEAUNE P. & BEIGBEDER M. (2013). Expansion sémantique des requêtes pour un modêle de recherche d'information par proximité. In *INFORSID*, p. 83–90.

BATES M. J. (1989). The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review*, **13**(5), 407–424.

BATES M. J. (1990). Where should the person stop and the information search interface start ? *Inf. Process. Manage.*, **26**(5), 575–591.

BRODER A. (2002). A taxonomy of Web search. *ACM SIGIR Forum*, **36**(2), 3–10.

CHAIGNAUD N., DELAVIGNE V., HOLZEM M., KOTOWICZ J.-P. & LOISEL A. (2010). Étude cognitive des processus de construction d'une requête dans un système de gestion de connaissances médicales. *Revue Technique et Science Informatiques (TSI)*, **29**, 991–1021.

DARMONI S., LEROY J., BAUDIC F., DOUYERE M., PIOT J. & THIRION B. (2000). CISMeF : a structured health resource guide. *Methods of Information in Medicine*, **39**, 30–35.

DUBUISSON DUPLESSIS G. (2014). *Modele de comportement communicatif conventionnel pour un agent en interaction avec des humains : Approche par jeux de dialogue*. PhD thesis, INSA de Rouen.

DUBUISSON DUPLESSIS G., CHAIGNAUD N., KOTOWICZ J., PAUCHET A. & PÉCUCHET J. (2013). Empirical specification of dialogue games for an interactive agent. In *Advances on Practical Applications of Agents and Multi-Agent Systems (PAAMS) Salamanca, Spain*, p. 49–60.

DUBUISSON DUPLESSIS G., PAUCHET A., CHAIGNAUD N. & KOTOWICZ J. (2015). A conventional dialogue model based on empirically specified dialogue games. In *nternational Conference on Tools with Artificial Intelligence (ICTAI), Vietri sul Mare, Italy*, p. 997–1004.

EL GUEDRIA Z. & VERCOUTER L. (2015). Personnalisation par un système multi-agent de la navigation au sein d'un corpus documentaire. In *JFSMA*, p. 61–70 : Cépaduès Éditions.

LOISEL A. (2008). *Modélisation du dialogue Homme-Machine pour la recherche d'informations : approche questions-réponses*. PhD thesis, INSA de Rouen.

MARCHIONINI G. (1989). Information-seeking strategies of novices using a full-text electronic encyclopedia. *Journal of the American Society for Information Science*, **40**(1), 54.

MARCHIONINI G. & WHITE R. (2007). Find What You Need, Understand What You Find. *Int. J. Hum. Comput. Interaction*, **23**(3), 205–237.

O'DAY V. L. & JEFFRIES R. (1993). Orienteering in an information landscape : how information seekers get from here to there. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, p. 438–445 : ACM.

SOUALMIA L. (2004). *Etude et évaluation d'approches multiples d'expansion de requêtes pour une recherche d'information intelligente : application au domaine de la santé sur l'Internet*. PhD thesis, INSA de Rouen.

SUTCLIFFE A. & ENNIS M. (1998). Towards a cognitive theory of information retrieval. *Interacting with computers*, **10**, 321–351.

Interprétation Interactive de connaissances à partir de traces

Amélie Cordier[1], Béatrice Fuchs[2].

[1] Université de Lyon, Claude Bernard, LIRIS, F-69 100, Villeurbanne, France, amelie.cordier@liris.cnrs.fr

[2] Université de Lyon, Jean Moulin, IAE, LIRIS, F-69 008, Lyon, France, beatrice.fuchs@liris.cnrs.fr

Résumé : Dans le processus d'extraction de connaissances à partir de données (ECD), l'analyste est au centre des opérations car c'est lui qui possède les connaissances pour interpréter les résultats. L'interactivité est alors déterminante lors de l'interprétation des motifs issus de la fouille pour choisir ceux qui deviendront des connaissances. Nous proposons une démarche d'interprétation interactive lors du processus d'extraction de connaissances à partir de données (ECD), dans le cadre de la recherche d'épisodes séquentiels à partir de traces. Elle s'appuie sur une visualisation des épisodes séquentiels obtenus par la fouille, sur lesquels l'analyste peut interagir. Il peut trier les résultats à l'aide de mesures de qualité, visualiser les occurrences de motifs sur la trace, et et un mécanisme de révision automatique permet de filtrer les motifs au voisinage d'un motif choisi. Des expérimentations montrent l'intérêt de cette approche qui est appliquée au domaine de l'analyse mélodique.

Mots-clés : extraction de connaissances, Visualisation, interactions, traces

1 Introduction

L'extraction de connaissances à partir de données (ECD) vise à découvrir des connaissances nouvelles dans de gros volumes de données à l'aide de méthodes non triviales, dans un processus *interactif* et *itératif* (Frawley *et al.*, 1992). La nature itérative du processus est due à la complexité du phénomène étudié dont chaque itération améliore graduellement la compréhension. L'interactivité quant à elle est due au travail d'un analyste, expert du domaine, qui joue un rôle central dans le processus d'ECD : il dirige le travail d'analyse en guidant les différentes étapes et en décidant quelles sont les connaissances pertinentes, celles qui font sens dans le domaine étudié. Sa présence est d'autant plus importante que les connaissances du domaine ne sont généralement pas disponibles dans le système (Mathern, 2012).

Les travaux dans le domaine de l'ECD se sont longtemps focalisés sur l'étape de fouille car elle est au centre du processus et pose des problématiques de calcul complexes. En effet, la fouille assure le traitement automatique de gros volumes de données pour y mettre en évidence des régularités. Mais pour passer des régularités aux connaissances, une expertise humaine est indispensable et le travail se heurte à plusieurs difficultés. D'une part le paramétrage de la fouille est loin d'être aisé et il faut en général s'y prendre à plusieurs fois pour trouver un paramétrage « convenable ». D'autre part l'étape d'interprétation est très délicate car des milliers de résultats doivent être traités manuellement. Cette dernière étape est pourtant déterminante, la forme des résultats doit être adéquate pour qu'ils soient compréhensibles, et leur présentation doit faciliter le travail de l'analyste afin de lui permettre de mobiliser ses connaissances. L'expérience a montré que l'utilisation de mesures d'intérêt indépendantes du domaines, des mesures d'intérêt *objectives*, telles que le support ou la longueur (Béchet *et al.*, 2014) sont insuffisantes lorsqu'il y a beaucoup de motifs et de redondance combinatoire (van Leeuwen, 2014). Par ailleurs, il y

Ce travail a été réalisé dans le cadre du projet OFS - Open Food System, programme investissements d'avenir.

a un réel besoin de suivre l'avancement du travail de l'analyste, de mémoriser le travail réalisé à chaque étape et de le capitaliser dans une perspective de réutilisation lors de sessions de découverte ultérieures liées à l'itérativité du processus.

On a donc rapidement pris conscience de l'attention à porter à toutes les étapes du processus d'ECD et plus particulièrement à leur caractère interactif qui est déterminant pour mener à bien le processus complet (Holzinger, 2013). Pour susciter la mobilisation cognitive de l'analyste, des outils sont nécessaires pour lui faciliter l'interprétation des résultats. Ces constats expliquent l'intérêt des travaux sur l'interactivité et la visualisation pour assister toutes les étapes du processus d'ECD (Kuntz *et al.*, 2006).

Nous proposons dans cet article une approche visant à assister le travail de l'analyste lors de l'interprétation en lui facilitant le suivi de l'avancement de son travail par une démarche itérative et interactive. A chaque itération, l'analyste peut visualiser et interagir sur les résultats de la fouille, et ses actions sont prises en compte pour gérer l'avancement de son travail et ainsi lui permettre de se focaliser plus rapidement sur des motifs d'intérêt. Dans la suite de l'article, nous situons ce travail dans le domaine de recherche, puis nous présentons le processus d'ECD et la démarche d'interprétation proposée associée aux définitions sous-jacentes, illustrées dans le domaine de l'analyse mélodique. Les premières expérimentations sont présentées ensuite afin étudier l'efficacité du processus, suivies d'une discussion. Pour finir, nous concluons sur l'état actuel du développement et ses perspectives.

2 Visualisation et interactions dans le processus d'ECD

Les représentation visuelles d'informations sont utilisées depuis longtemps en statistiques et en analyse de données et une communauté de chercheurs s'est constituée dans le domaine de la visualisation d'information. Avec l'émergence de la fouille de données et des algorithmes performants pour trouver des régularités dans de grandes quantités de données, on a rapidement pris conscience de l'enjeu à tirer parti des travaux sur la visualisation pour intégrer l'humain dans le processus de découverte de connaissances (Shneiderman, 2002), ce qui a abouti à la fouille visuelle des données qui s'est développée ces dernières années (Bertini & Lalanne, 2009). Plus généralement l'analyse visuelle (Keim *et al.*, 2010) vise à faire émerger des connaissances en combinant la puissance de traitement, la visualisation et l'expertise humaine, résumé par *"Analyse first, show the important, zoom, filter and analyse further, details on demand"*. Il s'agit donc de donner un rôle central et actif à l'humain dans le processus de découverte de connaissances (van Leeuwen, 2014). Ceci a donné lieu à plusieurs travaux dans ce sens avec une approche visuelle (Bothorel, 2014) et/ou interactive (Blanchard *et al.*, 2007a) pour les règles d'association.

Un des premiers problèmes de la fouille est la surabondance des résultats qui rend difficile leur exploration visuelle. Les travaux qui se sont intéressés à ce problème ont d'abord étudié des mesures d'intérêt objectives afin de caractériser la qualité des résultats de la fouille (Guillet & Hamilton, 2007), principalement les règles d'association. Puis les travaux ont visé à intégrer des connaissances du domaine dans le processus d'ECD, sous forme de bases de connaissances, d'ontologies ou de mesures subjectives (Marinica *et al.*, 2008; Brisson & Collard, 2008). La plupart des approches se sont intéressées aux règles d'association, ou aux règles temporelles (Blanchard *et al.*, 2007b, 2008), mais à notre connaissance, peu de travaux se sont intéressés aux épisodes séquentiels, aussi bien du point de vue des mesures d'intérêt que de la fouille visuelle. Par ailleurs, les interactions dans ces systèmes visent davantage à

changer de point de vue sur l'affichage les résultats, mais l'assistance à la construction d'un modèle est encore peu abordée dans la littérature (van Leeuwen, 2014).

Nous proposons un cadre pour la découverte d'épisodes séquentiels qui intègre un ensemble d'outils utiles pour l'analyse :

- un algorithme de fouille de séquences qui recherche des régularités dans des traces,
- une interface permettant des interactions avec l'analyste afin de visualiser une trace et des occurrences de motifs dans la trace,
- des mesures indépendantes du domaine pour caractériser la redondance combinatoire afin d'aider à identifier les motifs les plus prometteurs,
- un processus de révision qui consiste à filtrer les motifs au fur et à mesure des sélections de motifs afin d'assister le travail de l'analyste,
- un système à base de traces afin de mémoriser le résultat des analyses et capitaliser ainsi le travail réalisé pour des sessions de découverte ultérieures.

3 Extraction de connaissances à partir de traces

Les traces sont étudiées dans le cadre d'un processus classique d'ECD mis en œuvre dans un cycle composé des étapes principales de pré-traitement (sélection de trace, transformation), fouille, puis post-traitement (visualisation, interprétation). Bien que nous nous plaçons dans le cadre de l'étude des *traces*, les concepts présentés ici puissent s'appliquer à des données temporellement situées quelconques. Nous utilisons le domaine de l'analyse mélodique comme application « jouet » pour évaluer nos propositions. Il s'agit d'analyser une partition musicale décrite par une séquence de notes associées à une durée pour y détecter des motifs mélodiques récurrents.

3.1 Traces et système à base de traces

Notre approche s'appuie sur un système dédié à la gestion de traces qui permet d'une part de collecter des traces et les stocker, mais également à les manipuler à l'aide d'opérations génériques. Une trace est constituée d'une séquence d'éléments observés temporellement situés appelés des *obsels*. Elle est associée à un *modèle de trace* décrivant les types d'obsels, leurs attributs et leurs relations avec d'autres types d'obsels. Le modèle de trace permet d'interpréter les informations de la trace pour faciliter son exploitation ultérieure. Dans le domaine de l'analyse mélodique par exemple, les obsels décrivent les notes d'une partition musicale et sont caractérisés par un nom et une durée. Les traces sont manipulées par un ensemble d'opérations élémentaires appelées *transformations* qui sont de différents types : filtrage d'obsels, fusion de traces, etc. Parmi celles-ci, la *réécriture* crée une nouvelle trace appelée *trace transformée* qui vise à augmenter progressivement le niveau de compréhension et d'abstraction de la trace initiale. La réécriture consiste à construire une nouvelle trace t_2 à partir d'une trace primaire t_1 en remplaçant dans t_2 des motifs, c'est-à-dire des séquences d'obsels non nécessairement contigüs de t_1 par de nouveaux types d'obsels résumant chaque motif. Par exemple, la transformation $\bigstar$ est définie de la façon suivante : $\bigstar : \square\blacktriangle\triangledown\diamondsuit \longrightarrow \square\bigstar\triangledown$ avec $\bigstar = \blacktriangle\diamondsuit$. Un système à base de traces modélisées est un système permettant de collecter, de traiter et de visualiser des traces.

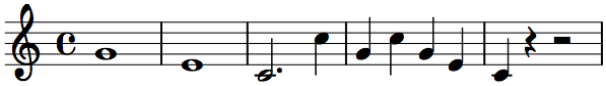
Le framework *kTBS* (*kernel for Trace Based System* [1]), (Champin *et al.*, 2013) réifie cette notion de système à base de traces. La réécriture de traces se situe au cœur du dispositif interactif mis en œuvre dans notre démarche d'interprétation.

3.2 Pré-traitement, fouille

Dans l'étape de pré-traitement, une trace est choisie par l'analyste dans une base de traces pour construire une séquence.

Définition 1 (Séquence)
Une séquence S est un ensemble d'événements typés et datés. Une occurrence d'événement est un couple (e_i, t_i) avec $e_i \in E$, où e_i est un type d'événement, E est l'ensemble des types d'événements et $t_i \in \mathbb{N}$ est l'estampille associée à e_i.

La séquence est construite à partir des obsels de la trace sélectionnée qui comportent une date de début et de fin et sont associés à un type d'obsel. Dans le cas de la partition musicale, les obsels sont des notes associées à une durée qui permettent de calculer leurs date de début et de fin, mais dans le cas général d'une trace, la durée n'est pas toujours disponible. Soit l'exemple de partition musicale suivant :



Les notes de cette partition peuvent être décrites par la trace suivante [2] :

Types d'obsels	G	E	C	C	G	C	G	E	C
durées	4	4	3	1	1	1	1	1	1
date de début	0	4	8	11	12	13	14	15	16

À partir de cette trace, la séquence suivante est construite avec $E = \{C, E, G\}$:
$S = \{(G, 0), (E, 4), (C, 8), (C, 11), (G, 12), (C, 13), (G, 14), (E, 15), (C, 16)\}$

À l'étape suivante, l'analyste fournit les paramètres de fouille qui servent à contraindre la fouille. L'étape de fouille utilise DMT4SP [3] (Nanni & Rigotti, 2007), un prototype d'extraction d'épisodes ou de règles séquentiels à partir d'une ou plusieurs séquences d'événements, conformément à la sémantique d'occurrence minimale (Mannila *et al.*, 1997). DMT4SP produit un ensemble de motifs fréquents satisfaisant les contraintes spécifiées dans le paramétrage.

Définition 2 (Motif, occurrence)
Un motif $m = (e_1, e_2, \ldots, e_n), e_i \in E$ est une séquence de types d'événements de longueur $l_m = n$. Une occurrence o_m^j du motif m est un ensemble d'estampilles $\{t_i\}_{i=1,n}$ tq $(e_i, t_i)_{i=1,n} \in S$. $O_m = \{o_m^j\}_j$ est l'ensemble des occurrences du motif m dans S.

Définition 3 (Fréquence, support)
On appelle fréquence ou support d'un motif m le nombre d'occurrence de ce motif dans S. Elle est notée $\sigma(m) = |O_m|$. La fouille retourne un ensemble M de motifs fréquents tels que $M = \{m_i\}, \forall i\ \sigma(m_i) \geqslant \sigma_{min}$, où σ_{min} est le support minimum choisi par l'analyste.

1. `http://tbs-platform.org/tbs/doku.php`
2. Ici les trois valeurs utilisées – la ronde, blanche pointée et noire ont pour durées respectives 4, 3 et 1 temps.
3. `http://liris.cnrs.fr/~crigotti/dmt4sp.html`

Dans l'exemple de partition musicale précédent, si $\sigma_{min} = 2$, les occurrences du motif (G, E, C) sont $o_1 = \{0, 4, 8\}$ et $o_2 = \{14, 15, 16\}$. $\sigma((G, E, C)) = 2 \geqslant \sigma_{min}$.

D'autres contraintes peuvent être spécifiées par l'analyste afin de limiter les résultats, comme la fenêtre temporelle, définie comme l'intervalle de temps maximal séparant le premier et le dernier événement des motifs [4].

3.3 Post-traitement et Interprétation interactive

Les motifs produits par la fouille sont mis en forme lors du post-traitement à l'aide des informations présentes dans la trace afin de les rendre intelligibles pour l'interprétation, puis ils sont affichés dans l'application TRANSMUTE. L'analyste peut trier les motifs selon plusieurs critères, les sélectionner, voir leurs occurrences dans la trace et choisir ceux qu'il estime les plus pertinents. L'interprétation consiste à construire une nouvelle trace transformée à partir de la trace analysée dans laquelle l'analyste mémorise les motifs qu'il a sélectionnés en créant de nouveaux types d'obsels qui se substituent aux occurrences des motifs sélectionnés dans la trace. Lorsqu'un motif est choisi, les occurrences des autres motifs ayant au moins un obsel en commun avec les occurrences du motif sélectionné sont éliminées, leur support est recalculé, et ceux dont le support est insuffisant sont éliminés : c'est l'opération de révision. La révision a pour effet de diminuer graduellement le nombre de résultats et facilite ainsi les prochains choix de l'analyste qui peut se focaliser sur d'autres motifs. Lorsque l'analyste a sélectionné tous les motifs et leur a associé les types d'obsels les remplaçant, il peut déclencher la réécriture qui procède à la création d'une nouvelle trace transformée. Ces opérations ainsi que le prototype TRANSMUTE sont présentés ci-après.

3.3.1 Le prototype Transmute

TRANSMUTE est un outil de génération de transformation de traces à partir des interactions de l'analyste qui permet d'afficher et d'interagir avec une trace et les motifs issus de la fouille. L'architecture de TRANSMUTE repose sur le module DISKIT qui met en œuvre le cycle d'ECD et SAMOTRACES, un framework Javascript pour la visualisation et les interactions (Barazzutti *et al.*, 2016). L'interface (figure 1) comporte dans la partie supérieure la trace en cours d'analyse associée à son modèle où les occurrences des motifs sélectionnés par l'analyste sont affichés et dans la partie inférieure, à gauche les paramètres de la fouille saisis par l'analyste et à droite les motifs obtenus par la fouille. Suite à la sélection de deux motifs, on observe dans la partie inférieure les motifs filtrés par la révision qui apparaîssent estompés dans l'interface.

3.3.2 Mesures d'intérêt

La fouille produit généralement un grand nombre de motifs caractérisés par une forte redondance combinatoire : un épisode apparaît sous la forme d'un grand nombre de variantes qui ne peuvent être éliminées faute de connaissances suffisantes, car elles satisfont toutes les contraintes qui ont été spécifiées. Le travail de l'analyste qui doit examiner ces résultats est fastidieux et le choix des « bons » épisodes est une tâche difficile. Afin d'aider l'analyste, des

4. Tous ces types de contraintes ne sont pas détaillés dans cet article.

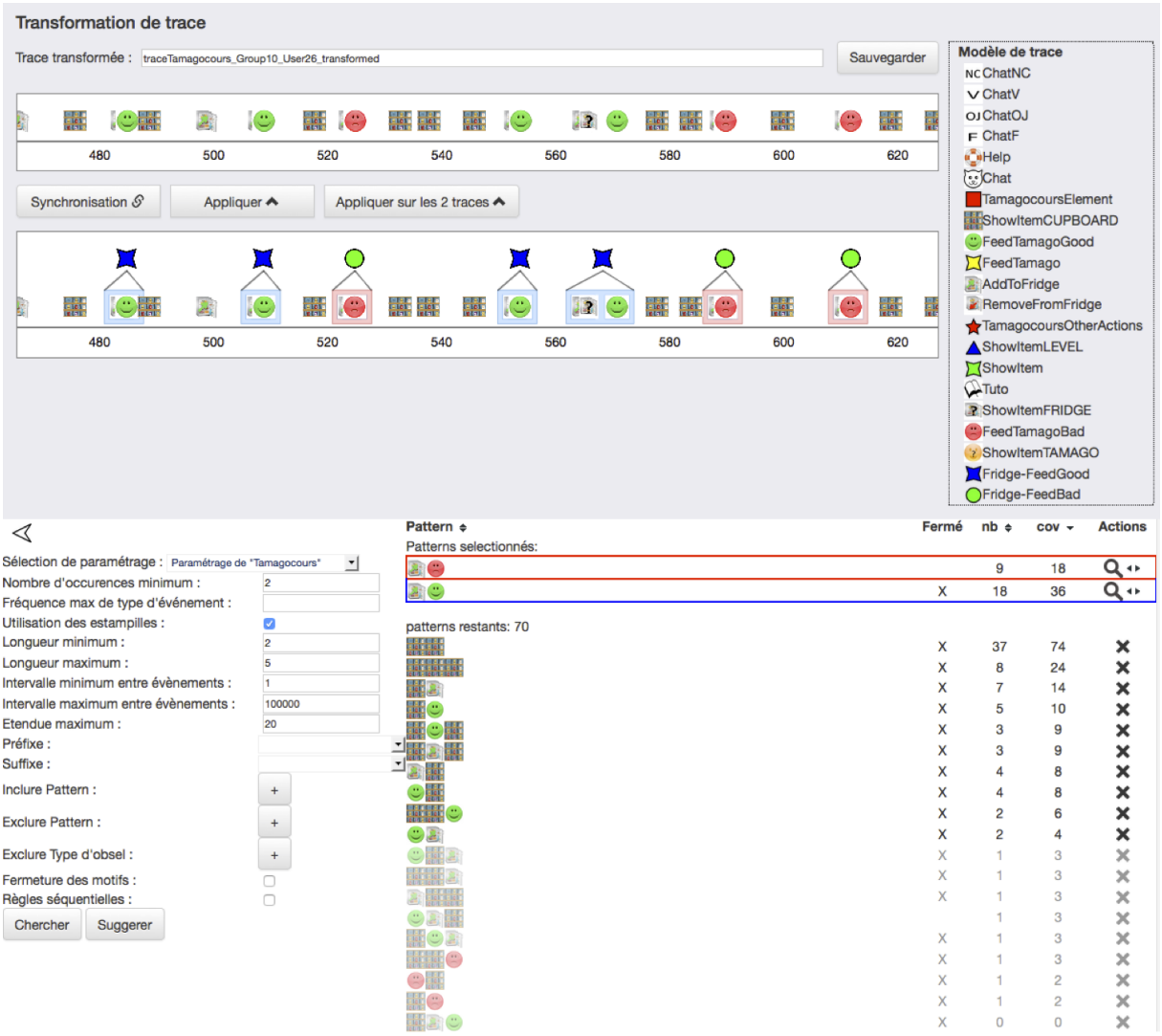


FIGURE 1 – L'interface de TRANSMUTE.

mesures sont utiles pour évaluer l'intérêt des motifs obtenus. Elles peuvent être spécifiques au domaine (subjectives), mais des mesures objectives peuvent également apporter une aide. En particulier nous nous sommes intéressés à des mesures objectives pour caractériser la redondance combinatoire des motifs.

Tout d'abord, la propriété de fermeture des motifs est importante à prendre en considération car elle fournit une représentation plus compacte des motifs et limite très fortement le nombre de motifs générés. Pour cette raison, de nombreux algorithmes de fouille la prennent en compte. Un épisode est fermé s'il n'est pas inclus dans un épisode de longueur supérieure ayant le même support [5]. Nous nous sommes limités à une seule séquence à l'aide de la fréquence telle que

5. Dans le cas des épisodes séquentiels, elle peut être définie de plusieurs façons lorsque l'on traite simultanément plusieurs séquences, car il existe plusieurs définitions du support : en terme de nombre d'occurrences d'un motif ou en terme de nombre de séquences où un motif est présent.

définie précédemment.

Définition 4 (Inclusion stricte de motifs)
Un motif $m_1 = (e_1, e_2, \dots e_p)$ *est inclus dans le motif* $m_2 = (e'_1, e'_2, \dots e'_n)$ *noté* $m_1 \prec m_2$ *si* $m_1 \neq m_2$ *et* $\exists \{j_i \in \mathbb{N}\}_{i=1\dots p}$ *tels que* $1 \leqslant j_i < j_{i+1} \leqslant n$ *et* $\forall i\ e_i = e'_{j_i}$

Définition 5 (Motif fermé)
Un motif m *est fermé si* $\forall\ m' \in M$ *tel que* $m \prec m'$ *alors* $\sigma(m) > \sigma(m')$

Dans l'exemple de la séquence
$S = \{(G,0),(E,4),(C,8),(C,11),(G,12),(C,13),(G,14),(E,15),(C,16)\}$,
le motif $m_1 = (G,E) \prec m_2 = (G,E,C)$ avec $\sigma(m_1) = \sigma(m_2)$, donc m_1 n'est pas fermé. En revanche, le motif $m_3 = (G,C) \prec m_2$ avec $\sigma((G,C)) = 3$, m_3 est fermé car il n'est inclus dans aucun motif ayant le même support.

En complement à la fermeture des motifs, nous avons défini les mesures ci-dessous pour évaluer la redondance combinatoire.

Définition 6 (Couverture événementielle)
La couverture événementielle CE_m *est l'ensemble des estampilles distinctes des occurrences d'un motif* m. $CE_m = \bigcup_{i=1}^{\sigma(m)} o_m^i$. *L'indicateur de couverture événementielle* IC_m *d'un motif est le nombre d'estampilles distinctes de ses occurrences :* $IC_m = |CE_m|$.

Définition 7 (Étalement événementiel)
L'étalement EE_m *d'un motif* m *de longueur* n *est le nombre d'événements de* S *dont l'estampille est incluse dans les intervalles des occurrences de* m. $EE_m = |\{t_k, \forall o_m^i \exists I_m^i, t_1^i \leqslant t_k \leqslant t_2^i\}|$

Définition 8 (Bruit)
Le bruit B_m *d'un motif est le nombre d'événements ne faisant partie des occurrences d'un motif donné et insérés dans l'intervalle temporel d'un motif :* $B_m = EE_m - CE_m$

Dans certains cas, il est possible de calculer des indicateurs temporels du même type que les indicateurs événementiels précédents mais qui portent sur la durée des événements, lorsque la durée des événements est disponible dans la trace. C'est le cas dans le domaine de l'analyse mélodique où les notes sont associées à une durée.

Définition 9 (Couverture temporelle)
La couverture temporelle CT_m *d'un motif* $m \in M$ *est la durée totale des événements de la couverture événementielle. On note* $d((e,t))$ *la durée associée à l'événement* (e,t). *Soit* m *un motif de couverture événementielle* $CE_m, \forall (e_i,t_i) \in CE_m, CT_m = \Sigma_i d((e_i,t_i))$

Dans l'exemple précédent, considérons les motifs $m_1 = (G,C)$ et $m_2 = (G,E,C)$.
$CE_{m_1} = \{0,8,12,13,14,16\}$, $IC_{m_1} = 6$, $EE_{m_1} = 8$, $B_{m_1} = 8-6 = 2$ et $CT_{m_1} = 11$.
$CE_{m_2} = \{0,4,8,14,15,16\}, IC_{m_2} = 6, EE_{m_2} = 6$, $B_{m_2} = 6-6 = 0$ et $CT_{m_2} = 14$.

Ces définitions permettent de décrire les principes de la révision interactive.

3.3.3 Révision interactive

Au fur et à mesure que des motifs sont choisis, il devient inutile de continuer à présenter tous les motifs qui sont, parfois à peu de choses près, des variantes des motifs choisis. Dans l'exemple de la partition précédente, si le motif (G, E, C) est choisi, il est inutile de considérer d'autres motifs tels que (G, E) ou (E, C) par exemple, car leur occurrences comportent des événements présents dans les occurrences de (G, E, C). Le choix d'un motif donné rend caduque l'examen de toutes ses variantes que l'analyste ne devrait pas avoir à considérer par la suite. Dans ce but, le principe de la révision repose sur le filtrage de tous les motifs redondants avec le motif choisi, afin de favoriser la focalisation sur d'autres motifs - ou régions de l'espace de recherche non encore explorés. L'étape d'interprétation est itérative et à chaque itération, le choix d'un motif est suivi d'une révision qui a pour conséquence un filtrage des motifs restant à considérer. La révision joue un rôle important dans l'assistance fournie à l'analyste en phase d'interprétation car elle agit comme un filtrage de l'espace de recherche autour d'un motif choisi par l'analyste. Elle se base sur la couverture événementielle pour rechercher les motifs à supprimer.

Définition 10 (Révision)
Soit m_c un motif choisi par l'expert de couverture événementielle CE_c.
Soit $m_i \in M, m_i \neq m_c$ un motif, l'ensemble des occurrences de m_i invalidées par le choix de m_c est : $O(m_i|m_c) = \{\forall o_i \in O(m_i), o_i \cap CE_c \neq \emptyset\}$.
On note $M_{m_c} \subset M$ l'ensemble des motifs de M invalidés par le choix de m_c par l'analyste : $M_{m_c} = \{m_i \in M, m_i \neq m_c,$ tq $\sigma(O(m_i|m_c)) < \sigma_{min}\}$
Lorsque le motif m_c a été choisi, l'ensemble des motifs restants à examiner par l'expert à l'itération suivante, est : $M \setminus M_{m_c}$.

Dans l'exemple, soient les motifs $m_1 = (G, E, C)$, $m_2 = (G, E)$ et $m_3 = (G, C)$.
$CE_{m_1} = \{0, 4, 8, 13, 15, 16\}$, les occurrences de m_2 sont $o^1_{m_2} = \{0, 4\}$ et $o^2_{m_2} = \{14, 15\}$, et pour m_3, $o^1_{m_3} = \{0, 8\}$,$o^2_{m_3} = \{12, 13\}$,$o^3_{m_3} = \{14, 16\}$. Si l'expert choisit le motif m_1, $o^1_{m_2}$ et $o^2_{m_2}$ sont supprimées ainsi que $o^1_{m_3}$ et $o^3_{m_3}$. Les supports de m_2 et m_3 deviennent respectivement 0 et $1 < \sigma_{min}$. La révision suite au choix de m_1 a donc pour effet d'éliminer m_2 et m_3.

La révision commence initialement avec l'ensemble M de tous les motifs issus de la fouille. A chaque itération, l'ensemble M est progressivement épuré de la redondance combinatoire autour d'un motif choisi en éliminant les motifs voisins dans l'espace de recherche.

Les mesures objectives ainsi que la révision interactive sont évalués dans le paragraphe suivant à l'aide d'une expérimentation dans le domaine de l'analyse musicale.

4 Expérimentations

Dans le domaine de l'analyse mélodique, trois partitions ont été étudiées pour lesquelles les motifs intéressants à retrouver dans les résultats de la fouille sont fournis par l'expert. Nous les appellerons dans la suite les *motifs experts*. Nous proposons de mesurer l'efficacité du processus par l'effort requis par l'analyste pour trouver les motifs intéressants en terme de nombre de motifs à examiner pour trouver tous les motifs experts, ce qui correspond au rang du dernier motif expert trouvé dans les résultats de fouille. Ainsi, plus ce rang est faible, plus l'effort requis par l'analyste est faible et plus la stratégie mise en place est efficace. L'effort de l'expert

est mesuré en triant les motifs selon plusieurs critères : l'ordre de sortie de la fouille (sans tri), la fréquence, l'indice de couverture événementielle, la couverture temporelle. Dans un deuxième temps le bruit est utilisé comme premier critère de tri et les quatre critères précédents comme deuxième critère. Ensuite, le rang le plus élevé des motifs experts est observé pour chaque critère de tri et indique l'effort requis pour trouver tous les motifs experts. Nous reportons les résultats dans deux tables ci-dessous : dans la première table, l'effort est d'abord mesuré sans révision, et dans la deuxième table, il est mesuré avec révision à chaque fois qu'un motif expert est trouvé. Intuitivement, la suppression de motifs à chaque fois qu'un motif expert est sélectionné devrait mener à une diminution du nombre de motifs et aura un effet sur les rangs des motifs experts restants qui devraient nécessairement diminuer. Néanmoins, des motifs experts sont susceptibles d'être éliminés à la suite de la révision, et il convient alors d'observer leur taux de rappel. Le paramétrage utilisé pour la fouille a été choisi de façon à assurer la présence de tous les motifs experts ($\sigma_{min} = 2$).

Sans révision, le rang du dernier motif expert est résumé dans le tableau 1 pour les trois pièces, selon le critère de tri utilisé ($\nearrow$ indique un tri croissant et $\searrow$ un tri décroissant).

pièce	motifs fouille	motifs experts	sans tri	$\sigma \searrow$	$IC \searrow$	$CT \searrow$	Bruit $\nearrow$ puis $\sigma \searrow$	Bruit $\nearrow$ puis $IC \searrow$	Bruit $\nearrow$ puis $CT \searrow$
1	3 853	11	3 838	3 838	3 797	2 735	240	369	**233**
2	12 947	20	12 818	12 829	12 591	**9 516**	12 667	12 672	12 668
3	59 786	29	53 805	56 061	31 309	30 151	**22 364**	25 317	25 420

TABLE 1 – Rang du dernier motif expert sans révision.

Sans révision, les meilleurs résultats sont obtenus avec un tri par bruit croissant et couverture temporelle décroissante pour la pièce 1, par couverture temporelle décroissante pour la pièce 2 et par bruit croissant puis fréquence décroissante pour la pièce 3, et un taux de rappel des motifs experts de 100% car aucun motif n'a été supprimé. La diminution de l'effort de l'expert est respectivement de 94%, 26% et 58% par rapport à un traitement sans tri. Dans tous les cas, l'utilisation des mesures a permis une diminution sensible de l'effort de l'expert.

Lorsque la révision est introduite, le rang du dernier motif expert trouvé est résumé dans le tableau 2. Les meilleurs résultats sont obtenus avec un tri par bruit croissant pour la pièce 1 et par couverture temporelle décroissante pour les pièces 2 et 3.

pièce	motifs fouille	motifs experts	sans tri	$\sigma \searrow$	$IC \searrow$	$CT \searrow$	Bruit $\nearrow$ puis $\sigma \searrow$	Bruit $\nearrow$ puis $IC \searrow$	Bruit $\nearrow$ puis $CT \searrow$
1	3 853	11	502	52	49	24	13	13	**12**
2	12 947	20	801	222	95	**71**	204	204	204
3	59 786	29	13490	5103	537	**527**	1533	1533	1533

TABLE 2 – Rang du dernier motif expert avec révision.

La diminution de l'effort est significative par rapport à un traitement sans tri et sans révision (tableau 3), et montre l'efficacité la révision conjointement aux mesures. Cependant, le rappel

des motifs experts est de 82% pour la pièce 1 et 100% pour la pièce 3 et pour la pièce 2, il est de 33% lorsque le tri n'utilise pas le bruit et de 62% lorsque le bruit est introduit. L'introduction du bruit a favorisé les motifs constitués de notes plus proches et amélioré le rappel, qui reste néanmoins encore décevant.

					Bruit ↗ puis		
pièce		$\sigma \searrow$	$IC \searrow$	$CT \searrow$	$\sigma \searrow$	$IC \searrow$	$CT \searrow$
1	Diminution	90%	90%	95%	97%	97%	98%
	Rappel	82%			82%		
2	Diminution	72%	88%	91%	75%		
	Rappel	33%			62%		
3	Diminution	62%	96%	96%	89%		
	Rappel	100%			100%		

TABLE 3 – Synthèse de la diminution de l'effort par rapport à un traitement des motifs de la fouille sans révision et sans tri et du taux de rappel des motifs experts.

L'ordre dans lequel les motifs sont choisis a une incidence sur les motifs éliminés par la révision, d'où l'importance du choix des critères de tri. Ces résultats montrent qu'il est important de disposer de plusieurs mesures pour tenir compte des caractéristiques de chaque pièce qu'il convient par conséquent d'observer avant l'analyse pour prendre en compte leurs particularités. Des mesures subjectives telle que celle présentée dans (Fuchs, 2011), n'ont pas été introduites dans ce travail et sont indispensables pour compléter ces mesures objectives. On peut également mentionner que TRANSMUTE permet d'annuler une sélection de motif, et que cette souplesse d'utilisation n'a pas été prise en compte dans cette expérimentation.

5 Discussion

Le prototype Transmute qui met en oeuvre cette approche possède des limitations principalement liées à l'interface, car il n'a pour l'instant bénéficié d'aucune optimisation. Il est utilisable sur de petites traces, et ne permet pas de traiter un nombre de motifs trop important (quelques milliers d'obsels et de motifs), et il reste du travail à réaliser pour lever ce verrou. Le module DISKIT en revanche peut traiter des traces et un nombre de motifs beaucoup plus importants. Une validation qualitative a montré que TRANSMUTE a pu être pris en main aisément par les utilisateurs (Barazzutti, 2015). TRANSMUTE s'est également avéré utilisable pour l'analyse de traces d'un jeu sérieux collaboratif pour l'apprentissage de règles de diffusion de ressources numériques (TAMAGOCOURS) montré à la figure 1. Une base de trace a été construite à partir des sessions de jeu de 244 étudiants répartis en 86 groupes représentant au total environ 26 000 obsels. Les groupes et utilisateurs ont été analysés séparément, représentant des traces de quelques centaines d'obsels.

Actuellement, les mesures permettant de trier les motifs sont prédéfinies dans TRANSMUTE. Il serait souhaitable que l'analyste puisse choisir lui-même des mesures. Des mesures subjectives peuvent être conçues à l'aide du kTBS, car les motifs issus de la fouille sont mis en relation avec les obsels de la trace analysée. Leurs attributs et relations rendent possible la réalisation de

calculs plus complexes qui ne peuvent bien-entendu être implémentés que par un informaticien et mis à disposition d'un analyste.

6 Conclusion et perspectives

Nous avons présenté une démarche d'interprétation itérative et interactive dans un processus d'ECD à partir de traces. Elle s'appuie sur l'utilisation de mesures d'intérêt pour trier les motifs, une visualisation des motifs issus de la fouille où l'analyste peut interagir pour voir l'impact de ses actions sur la trace. Un filtrage dynamique au voisinage des motifs sélectionnés favorise une meilleure focalisation sur de nouvelles régions de l'espace de recherche. La création d'une trace transformée et la réécriture permettent de mémoriser le travail de l'analyste afin de le prendre en compte lors de sessions de travail ultérieures. Les premières expérimentations réalisées sont encourageantes. La plateforme TRANSMUTE réifie cette approche.

Outre l'approfondissement des mesures pour l'interprétation, une piste d'amélioration consiste à explorer dans leur ensemble les motifs pour sélectionner ceux qui présentent ensemble une meilleure couverture globale de la séquence initiale (Vreeken *et al.*, 2010), ou encore de traiter les motifs par groupes selon une mesure de similarité. D'autres perspectives concernent l'assistance aux autres phases du processus d'ECD et en particulier le pré-traitement. Tout d'abord un « bon » paramétrage de la fouille n'est pas une tâche facile et nous pensons poursuivre le travail sur l'interactivité pour aider l'analyste dans la phase de paramétrage. Pour cela, nous considérons plusieurs pistes : (1) un processus d'ECD entièrement interactif où les interactions avec l'analyste et les résultats de la fouille sont utilisés pour guider le réglage des paramètres, et (2) la recommandation de paramétrages à partir d'expériences antérieures ou à partir des interactions de l'utilisateur sur la trace elle-même. Plus généralement, la préparation des données peut être enrichie avec toutes sortes de requêtes sur les traces afin de varier les dimensions à analyser. De plus, la prise en compte de plusieurs dimensions permettrait une analyse plus fine et précise dans le domaine de l'analyse musicale. Enfin l'analyse simultanée de plusieurs traces et la façon aborder la représentation des motifs sur un grand nombre de traces est un sujet important dès lors qu'il s'agit d'étudier par exemple les traces d'utilisateurs différents ou des pièces musicales à plusieurs voix.

Références

BARAZZUTTI P.-L. (2015). *Transmute : un outil interactif d'assistance à la découverte de connaissances*. Mémoire de master en informatique, Université Claude Bernard Lyon 1.

BARAZZUTTI P.-L., CORDIER A. & FUCHS B. (2016). Transmute : un outil interactif pour assister l'extraction de connaissances à partir de traces. In B. CRÉMILLEUX & C. DE RUNZ, Eds., *Extraction et Gestion des Connaissances - EGC 2016*, volume RNTI-E-30 of *Extraction et Gestion des Connaissances*, p. 463–468, Reims, France : Cyril de Runz RNTI.

BÉCHET N., CELLIER P., CHARNOIS T. & CRÉMILLEUX B. (2014). Fouille de motifs séquentiels pour la découverte de relations entre gènes et maladies rares. *Revue d'Intelligence Artificielle*, **28**(2-3), 245–270.

BERTINI E. & LALANNE D. (2009). Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. In *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery : Integrating Automated Analysis with Interactive Exploration*, p. 12–20 : ACM.

BLANCHARD J., GUILLET F. & BRIAND H. (2007a). Interactive visual exploration of association rules with rule-focusing methodology. *Knowledge and Information Systems*, **13**(1), 43–75.

BLANCHARD J., GUILLET F. & GRAS R. (2007b). On the discovery of significant temporal rules. In *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*, p. 443–450 : IEEE.

BLANCHARD J., GUILLET F. & GRAS R. (2008). Assessing the interestingness of temporal rules with sequential implication intensity. In *Statistical Implicative Analysis*, p. 55–71. Springer.

BOTHOREL G. (2014). *Algorithmes automatiques pour la fouille visuelle de données et la visualisation de règles d'association : application aux données aéronautiques*. PhD thesis.

BRISSON L. & COLLARD M. (2008). How to semantically enhance a data mining process ?. In *ICEIS*, volume 19, p. 103–116 : Springer.

CHAMPIN P.-A., MILLE A. & PRIÉ Y. (2013). Vers des traces numériques comme objets informatiques de premier niveau : une approche par les traces modélisées. *Intellectica*, (59), 171–204.

FRAWLEY W. J., PIATETSKY-SHAPIRO G. & MATHEUS C. J. (1992). Knowledge discovery in databases : An overview. *AI Magazine*, **13**(3), 57–70.

FUCHS B. (2011). Co-construction interactive de connaissances. Application à l'analyse mélodique. In A. MILLE, Ed., *Ingénierie des connaissances*, p. 705–720.

GUILLET F. & HAMILTON H. J. (2007). *Quality measures in data mining*, volume 43. Springer.

HOLZINGER A. (2013). Human-computer interaction and knowledge discovery (hci-kdd) : What is the benefit of bringing those two fields to work together ? In *Availability, Reliability, and security in Information Systems and HCI*, p. 319–328. Springer.

KEIM D. A., KOHLHAMMER J., ELLIS G. & MANSMANN F. (2010). *Mastering the information age-solving problems with visual analytics*. Florian Mansmann.

KUNTZ P., LEHN R., GUILLET F. & PINAUD B. (2006). Découverte interactive de règles d'association via une interface visuelle. *Visualisation en Extraction des Connaissances*, p. 113–125.

MANNILA H., TOIVONEN H. & VERKAMO A. I. (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, **1**, 259–289.

MARINICA C., GUILLET F. & BRIAND H. (2008). Post-processing of discovered association rules using ontologies. In *Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on*, p. 126–133 : IEEE.

MATHERN B. (2012). *Découverte interactive de connaissances à partir de traces d'activité : Synthèse d'automates pour l'analyse et la modélisation de l'activité de conduite automobile*. Thèse de doctorat en informatique, Université Claude Bernard Lyon 1.

NANNI M. & RIGOTTI C. (2007). Extracting trees of quantitative serial episodes. In S. DŽEROSKI & J. STRUYF, Eds., *Knowledge Discovery in Inductive Databases : 5th International Workshop, KDID 2006 Berlin, Germany, September 18, 2006 Revised Selected and Invited Papers*, p. 170–188, Berlin, Heidelberg : Springer Berlin Heidelberg.

SHNEIDERMAN B. (2002). Inventing discovery tools : combining information visualization with data mining. *Information visualization*, **1**(1), 5–12.

VAN LEEUWEN M. (2014). Interactive data exploration using pattern mining. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, p. 169–182. Springer.

VREEKEN J., LEEUWEN M. & SIEBES A. (2010). Krimp : mining itemsets that compress. *Data Mining and Knowledge Discovery*, **23**(1), 169–214.

Informations personnelles, Communautés et Recommandation

Étude du profil utilisateur pour la recommandation dans les folksonomies

Mohamed Nader Jelassi[1,2,3], Sadok Ben Yahia[1] et Engelbert Mephu Nguifo[2,3]

[1] Université Tunis El Manar. Faculté des Sciences de Tunis, Tunis, Tunisie.

[2] Clermont Université, Université Blaise Pascal, LIMOS, BP 10448, F-63000 Clermont-Ferrand, France.

[3] CNRS, UMR 6158, LIMOS, F-63171 Aubière, France.
{nader.jelassi@isima.fr, sadok.benyahia@fst.rnu.tn, engelbert.mephu_nguifo@univ-bpclermont.fr}

Résumé : Dans les *folksonomies*, les utilisateurs partagent des ressources (films, livres, sites web, etc.) en les annotant avec des tags librement choisis. Dans ce papier, nous considérons une nouvelle dimension dans une *folksonomie* qui contient des des informations supplémentaires sur les utilisateurs. Nous définissons un degré de proximité entre deux utilisateurs comme le nombre d'informations de profil en commun entre eux et nous proposons un système personnalisé de recommandations basé sur cette définition. Les expérimentations menées sur un jeu de données du monde réel, MOVIELENS, montrent l'utilité de la nouvelle dimension introduite et quelles informations sont les plus influentes durant le processus de recommandation.

Mots-clés : Folksonomie, Recommandation, Qualité, Précision, Informations supplémentaires, Profil

1 Introduction et Motivations

Une *folksonomie* désigne un système de classification collaborative par les internautes [1] [2] (Mika (2007)). Elle est composée de trois ensembles : un ensemble $\mathcal{U}$ d'utilisateurs, un ensemble $\mathcal{T}$ de tags (ou mots-clés) et un ensemble $\mathcal{R}$ de ressources (films, livres, sites web, photos, etc.) ((Hotho *et al.*, 2006)). Les utilisateurs sont les responsables du partage des ressources et l'affectation de tags à ces derniers (Strohmaier *et al.* (2012)). Cependant, il s'avère que le choix de tags et de ressources partagés par un utilisateur d'une *folksonomie* varie selon plusieurs critères : le genre, l'âge ou encore la profession de celui qui partage l'information. Cela a motivé les chercheurs à proposer des systèmes de recommandation personnalisés afin de répondre aux besoins de chaque utilisateur selon son profil. Ainsi, un système de recommandation personnalisé offre à l'utilisateur des tags et ressources en respectant le profil de ce dernier de telle sorte que les recommandations soient le plus proche de ses besoins (Ricci *et al.* (2011)) ((Nanopoulos *et al.*, 2010)). Pour répondre à cette tâche, nous considérons, dans ce papier, une quatrième dimension dans une *folksonomie*. Cette quatrième dimension peut recouvrir différents aspects : par exemple le profil (genre, âge, profession, ...) comme mentionné ci-dessus, ou le temps si on veut étudier la dynamique temporelle des *folksonomies*. Dans ce papier, nous traitons la quatrième dimension de manière indifférente pour l'aspect méthodologique, mais afin d'avoir des informations disponibles dans un jeu de données du monde réel et d'étudier l'influence de ces informations, nous focaliserons sur l'aspect profil. Par ailleurs, nous définissons un degré de proximité entre deux utilisateurs correspondant au nombre de

1. `http://www.vanderwal.net/folksonomy.html`

2. `http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html`

leurs variables (informations de profil, date de partage, etc.) en commun. Ainsi, grâce au système personnalisé de recommandation, que nous introduisons, qui repose sur cette mesure, nous montrons l'utilité des variables introduites dans la quatrième dimension durant le processus de recommandation. Enfin, nous menons des expérimentations sur un jeu de données du monde réel, *i.e.,* MOVIELENS, pour savoir quelles informations sont les plus influentes pour la recommandation. De plus, nous comparons les précisions de notre système de recommandation, avec et sans considération d'informations supplémentaires, aux approches de la littérature, et nous répondons aux questions suivantes : *(i)* est-ce que la dimension introduite est une information importante aidant à améliorer les recommandations dans les *folksonomies* ? ; *(ii)* à quel point la nouvelle dimension peut-être utile pour fournir des recommandations avec une meilleure précision ? ; *(iii)* et quelles informations supplémentaires sont les plus influentes durant le processus de recommandation ?

Le reste du papier est organisé comme suit : dans la section suivante, nous étudions les principales approches de la littérature. Nous présentons les notions de base dans la Section 3. Ensuite, dans la Section 4, nous introduisons notre système personnalisé de recommandation. Dans la Section 5, nous discutons les résultats de notre étude expérimentale. Enfin, nous concluons notre papier avec des perspectives pour nos travaux futurs dans la Section 6.

2 Travaux connexes

Dans un souci d'améliorer les recommandations dans les *folksonomies*, plusieurs travaux ont été proposés dans la littérature ((De Meo *et al.*, 2010)) ((Basile *et al.*, 2007)) ((Liang *et al.*, 2010)). Dans (Diederich & Iofciu (2006)), les auteurs utilisent la *"personomie"* d'un utilisateur, *i.e.,* les tags qui lui sont relatifs, afin de lui recommander des utilisateurs ayant partagé des tags et ressources similaires. Tout d'abord, ils construisent un profil pour chaque utilisateur. Ensuite, à partir de ce profil, les auteurs sont capables de recommander des utilisateurs (dits *collaborateurs*) en utilisant une mesure de similarité entre utilisateurs. Cette mesure, qui s'appuie uniquement sur les tags utilisés par les utilisateurs, n'offre pas une information complète sur les utilisateurs. Plus récemment, dans (Hu *et al.* (2011)), les auteurs se basent à la fois sur l'historique de tagging (tags et ressources) des utilisateurs et sur leurs contacts sociaux. La limite de cette approche est qu'elle requiert qu'un utilisateur doit posséder des contacts sociaux afin d'avoir des recommandations de tags. Dans (Jäschke *et al.* (2007)), Hotho *et al.* ont proposé des recommandations de tags dans les *folksonomies* basées sur les tags les plus utilisés. Cependant, ces recommandations ne sont absolument pas personnalisées étant donné que les mêmes tags sont proposés à chaque utilisateur. Lipczak a proposé dans (Lipczak (2008)) un système de recommandation de tags en trois étapes. À partir des tags annotés aux ressources, l'auteur ajoute des tags proposés par un lexique basé sur les co-occurrences de tags sur les mêmes ressources. Ensuite, le système filtre les tags déjà utilisés par l'utilisateur. Toutefois, malgré cette étape de filtrage, la recommandation ne paraît pas être personnalisée étant donné qu'elle cherche des tags co-occurrant sur d'autres annotations. L'approche revient ensuite à enlever les tags précédemment annotés par l'utilisateur de ceux qui sont suggérés. Dans (Landia & Anand (2009)), les auteurs ont proposé une nouvelle approche combinant la similarité à la fois entre ressources et entre utilisateurs afin de recommander des tags personnalisés. En effet, deux utilisateurs sont considérés comme similaires s'ils ont assigné les mêmes tags aux mêmes ressources. Toutefois, il est rare de trouver pareille situation dans des *folksonomies* où les tags utilisés par deux

utilisateurs sur les mêmes ressources sont identiques.

Dans notre approche, nous insistons sur le nécessaire recours à des informations supplémentaires et à les combiner à l'historique de tagging afin d'améliorer les recommandations. Toutes ces informations seront représentées par des quadri-concepts. Ainsi, dans ces structures, nous nous focalisons non seulement sur les tags/ressources les plus utilisés, mais également sur ceux qui ont été utilisés en combinaison par des utilisateurs *proches*, obtenant ainsi un résultat plus spécifique. Contrairement aux approches de la littérature qui se limitent à l'information <utilisateur, tag, ressource>, nous étendons ce triplet par l'information contenue dans la quatrième dimension. De plus, les quadri-concepts sont une représentation condensée, sans perte d'information, d'une *folksonomie* dont les données sont souvent altérées par des tags redondants ou par des utilisateurs inactifs.

Dans ce qui suit, nous présentons quelques notions qui seront utilisées tout au long de ce papier.

3 Notions de base

Nous commençons par présenter une extension de la notion de *folksonomie* (Jäschke *et al.* (2008)) par l'ajout d'une quatrième dimension ((Jelassi *et al.*, 2015)).

Définition 1
Une ***v-folksonomie*** *est un ensemble de tuples* $\mathcal{F}_v = (\mathcal{U}, \mathcal{T}, \mathcal{R}, \mathcal{V}, \mathrm{Y})$ *où* $\mathcal{U}$, $\mathcal{T}$, $\mathcal{R}$ *et* $\mathcal{V}$ *sont des ensembles finis dont les éléments sont appelés* ***utilisateurs***, ***tags***, ***ressources*** *et* ***variables****.* $\mathrm{Y} \subseteq \mathcal{U} \times \mathcal{T} \times \mathcal{R} \times \mathcal{V}$ *représente une relation quadratique où chaque élément* $y \subseteq \mathrm{Y}$ *peut être représenté par un quadruplet :* $y = \{(u, t, r, v) \mid u \in \mathcal{U}, t \in \mathcal{T}, r \in \mathcal{R}, v \in \mathcal{V}\}$ *ce qui veut dire que l'utilisateur* u *a annoté la ressource* r *via le tag* t *à travers la variable* v*. Nous considérons que deux utilisateurs sont* proches *s'ils partagent au moins une même variable en commun.*

La quatrième dimension introduite peut recouvrir différents aspects : par exemple le profil (genre, âge, profession, ...), ou le temps si on veut étudier la dynamique temporelle des *folksonomies*. Ainsi, l'information inclue dans la quatrième dimension est complètement corrélée au triplet (utilisateur, tag, ressource). Par exemple, dans un quadruplet (u, t, r, *date*), l'information *date* est corrélée à l'opération de tagging faite par u avec le tag t sur la ressource r. Dans un autre quadruplet (u, t, r, *profil*), l'information *profil* est corrélée aussi bien à l'utilisateur u qu'au tag t et la ressource r partagés par u. En effet, un utilisateur u peut partager un livre sur les langages de programmation avec le tag *programming* via le profil p_1 (*étudiant* par exemple) tandis qu'il peut partager un papier d'une revue scientifique avec le tag *paper* via le profil p_2 (*chercheur* par exemple). Dans ce papier, nous traitons la quatrième dimension de manière indifférente pour l'aspect méthodologique, mais afin d'avoir des informations disponibles dans un jeu de données du monde réel, nous focaliserons sur l'aspect profil comme quatrième dimension de la *folksonomie*.

Exemple 1
Le Tableau 1 montre un exemple d'une v-folksonomie $\mathcal{F}_v$ *avec* $\mathcal{U} = \{u_1, u_2, u_3, u_4\}$, $\mathcal{T} = \{t_1, t_2, t_3, t_4\}$, $\mathcal{R} = \{r_1, r_2, r_3\}$ *et* $\mathcal{V} = \{v_1, v_2\}$*. Chaque croix d'une relation quadratique, indique une opération de tagging faite par un utilisateur de l'ensemble* $\mathcal{U}$ *avec une variable de* $\mathcal{V}$*, utilisant*

un tag de $\mathcal{T}$ sur une ressource de $\mathcal{R}$. Par exemple, l'utilisateur u_1, qui a entre 25 et 35 ans et qui est étudiant, a taggué les articles r_1, r_2 et r_3 avec les tags thesis, web et to_recommend.

$\mathcal{F}_v$	$\mathcal{R}$	r_1				r_2				r_3			
$\mathcal{V}$	$\mathcal{U}/\mathcal{T}$	t_1	t_2	t_3	t_4	t_1	t_2	t_3	t_4	t_1	t_2	t_3	t_4
	u_1		×	×	×		×	×	×		×	×	×
v_1	u_2		×	×	×	×	×	×	×	×	×	×	×
	u_3		×	×	×	×	×	×	×	×	×	×	×
	u_4		×	×		×			×	×			×
	u_1		×	×	×		×	×	×		×	×	×
v_2	u_2		×	×	×	×			×	×	×	×	×
	u_3												
	u_4												

TABLE 1 – Un exemple d'une *v-folksonomie* avec les valeurs suivantes : *science*(t_1), *thesis*(t_2), *web*(t_3), *to_recommend*(t_4), *25-35 ans*(v_1), *étudiant*(v_2) et r_1, r_2 et r_3 trois articles scientifiques.

Nous définissons maintenant un concept quadratique (Jelassi *et al.* (2015)).

Définition 2
*Un concept quadratique (ou quadri-concept) d'une v-folksonomie $\mathcal{F}_v = (\mathcal{U}, \mathcal{T}, \mathcal{R}, \mathcal{V}, \mathrm{Y})$ est un quadruplet (U, T, R, V) avec $U \subseteq \mathcal{U}$, $T \subseteq \mathcal{T}$, $R \subseteq \mathcal{R}$ et $V \subseteq \mathcal{V}$ avec $U \times T \times R \times V \subseteq \mathrm{Y}$ tel que le quadruplet (U, T, R, V) est maximal, i.e., aucun de ces ensembles ne peut être augmenté sans diminuer un des trois autres ensembles. Pour un quadri-concept $QC = (U, T, R, V)$, U, R, T et V sont, respectivement, appelés **Extent**, **Intent**,* **Modus** *et **Variable**.*

Remarque 1
Afin de permettre l'extraction de l'ensemble de quadri-concepts fréquents à partir d'une v-folksonomie *donnée, nous pouvons utiliser l'un des deux algorithmes de la littérature dédiés à cette tâche :* QUADRICONS *(Jelassi* et al. *(2013)) ou* DATAPEELER *(Cerf* et al. *(2009)). Les deux algorithmes prennent en entrée une* v-folksonomie *ainsi que quatre seuils minimaux de support (un pour chaque dimension) et donnent en sortie l'ensemble de **quadri-concepts** vérifiant ces seuils. Un quadri-concept **fréquent** est un quadri-concept dont chaque ensemble (utilisateur, tag, ressource et variable) a une cardinalité supérieure ou égale à son seuil de support correspondant.*

Définition 3
*(*DEGRÉ DE PROXIMITÉ*) Considérons une v-folksonomie $\mathcal{F}_v = (\mathcal{U}, \mathcal{T}, \mathcal{R}, \mathcal{V}, \mathrm{Y})$, nous définissons le **degré de proximité** entre deux utilisateurs de $\mathcal{U}$ comme le nombre de leurs variables de $\mathcal{V}$ en commun.*

Exemple 2
Considérons la v-folksonomie, *représentée par le Tableau 1, le quadri-concept ($\{u_1, u_2, u_3\}$, $\{t_2, t_3, t_4\}$, $\{r_1, r_2, r_3\}$, v_1) montre que les utilisateurs u_1, u_2 et u_3 ont un degré de proximité égal à 1,* i.e., *ils partagent la variable v_1 en commun. Par contre, le quadri-concept ($\{u_1, u_2\}$, $\{t_2$,*

t_3, t_4}, {r_1, r_3}, {v_1, v_2}) *montre que les utilisateurs* u_1 *et* u_2 *ont deux variables en commun,* i.e., v_1 *et* v_2. *Ainsi, ils ont un degré de proximité égal à* 2.

4 Recommender : un nouvel algorithme pour des recommandations personnalisées

Dans cette section, nous proposons notre nouveau système personnalisé de recommandation RECOMMENDER. Le pseudo code de RECOMMENDER est présenté par l'Algorithme 1. RECOMMENDER prend en entrée $\mathcal{QC}$ (un ensemble de quadri-concepts fréquents) ainsi qu'un utilisateur cible u avec son ensemble de variables V, un degré de proximité d et (optionellement) une ressource r (à annoter). L'ensemble $\mathcal{QC}$ est extrait dans une étape de pré-traitement par l'un des algorithmes de la littérature dédiés à cette tâche. Ensuite, RECOMMENDER donne en sortie trois ensembles : un ensemble d'utilisateurs proposés, un ensemble de tags suggérés et un ensemble de ressources recommandées en prenant en considération le degré de proximité d entre utilisateurs. Ce degré est une métrique définie par l'utilisateur et qui est égale, au minimum, à 0 et, au maximum, aux nombre de variables disponibles dans le jeu de données considéré.

RECOMMENDER opère comme suit : il commence par initialiser tous les ensembles de sortie aux ensembles nuls (Ligne 2). Ensuite, il extrait les tags et ressources déjà partagés par l'utilisateur u (Lignes 3-4) afin d'éviter de lui recommander des tags et des ressources qu'il a déjà partagé. Par suite, selon le degré de proximité, RECOMMENDER opère comme suit : (*i*) si la valeur de d est égale à 0 (Lignes 5-12), *i.e.,* la recommandation est indépendante de l'ensemble de variables V, alors, nous recommandons à u un ensemble d'utilisateurs $\mathcal{PU}$ qui ont partagé les mêmes tags et ressources que lui (Ligne 9). De plus, nous recommandons à u un ensemble de ressources partagés par les utilisateurs de l'ensemble $\mathcal{PU}$ (Ligne 10). Enfin, nous recommandons également à u un ensemble de tags utilisés par ces mêmes utilisateurs sur la ressource ra que u souhaite partager (Lignes 11 et 12) ; (*ii*) si la valeur du degré de proximité est égal, au moins, à 1, *i.e.,* l'utilisateur u doit partager, au moins, d variables en commun avec les autres utilisateurs de la *v-folksonomie* (Ligne 15). En parcourant les quadri-concepts de l'ensemble $\mathcal{QC}$, si u appartient déjà à un quadri-concept qc, alors qc est élagué (Ligne 16) afin de filtrer les tags et ressources déjà partagés par u. Cette stratégie est inspirée par celle de (Lipczak (2008)). Ensuite, selon la tâche à accomplir, RECOMMENDER fonctionne comme suit : pour la tâche de *Proposition d'utilisateurs* (Ligne 7), c'est la partie *utilisateurs* du quadri-concept qc qui est ajoutée à l'ensemble $\mathcal{PU}$ des utilisateurs proposés. Cette tâche aide à connecter les utilisateurs qui ont des intérêts communs et aide également à promouvoir le partage de ressources. Pour la tâche de *Suggestion de tags* (Lignes 19 et 20), le but est de suggérer des tags personnalisés à un utilisateur qui souhaite ajouter une ressource à la *folksonomie*. Cette tâche a plusieurs avantages : elle rappelle à l'utilisateur ce dont une ressource s'agit, accroît l'annotation des ressources et permet de consolider le vocabulaire des utilisateurs (Ricci *et al.* (2011)). Pour cette tâche, nous ajoutons donc les tags affectés à la ressource ra par les utilisateurs qui ont d variables en commun avec u à l'ensemble $\mathcal{ST}$. Quant à la tâche de *Recommandation de ressources* (Ligne 18), le but est de proposer une liste personnalisée de ressources conforme aux intérêts de l'utilisateur u ; ces ressources sont ajoutées à l'ensemble $\mathcal{RR}$.

Algorithme 1 : RECOMMENDER

Données :

1. $\mathcal{QC}$: un ensemble de quadri-concepts fréquents
2. u : un utilisateur cible avec son ensemble de variables V
3. d : un degré de proximité
4. ra : une resource à annoter par u

Résultats :

1. $\mathcal{PU}$: un ensemble d'utilisateurs proposés
2. $\mathcal{ST}$: un ensemble de tags suggérés
3. $\mathcal{RR}$: un ensemble de ressources recommandées

```
1  début
2      PU=ST=RR=∅
3      u.Tags={t ∈ T / ∃ r ∈ R ∃ u ∈ U ∃ v ∈ V, (u,t,r,v) est un quadri-concept} ;
4      u.Ressources={r ∈ R / ∃ t ∈ T ∃ r ∈ R ∃ v ∈ V, (u,t,r,v) est un quadri-concept} ;
5      si d=0 alors
6          pour chaque quadri-concept qc ∈ QC faire
7              si u ∈ qc.Extent alors
8                  pour chaque utilisateur u' de qc.Extent faire
9                      PU = PU ∪ u' /*Proposition d'utilisateurs*/
10                     RR = RR ∪ u'.Resources \ u.Resources ; /*Recommandation de
                       ressources*/
11                     si ra ∈ qc.Intent alors
12                         ST = ST ∪ u'.Tags \ u.Tags ; /*Suggestion de tags*/
13     sinon si d > 0 alors
14         pour chaque quadri-concept qc ∈ QC faire
15             si | V ∩ qc.Variable | ≥ d alors
16                 si u ∉ qc.Extent alors
17                     PU = PU ∪ qc.extent /*Proposition d'utilisateurs*/
18                     RR = RR ∪ qc.intent \ u.Resources ; /*Recommandation de
                       ressources*/
19                     si ra ∈ qc.Intent alors
20                         ST = ST ∪ qc.modus \ u.Tags ; /*Suggestion de tags*/
21     retourner (PU,ST,RR) ;
22 fin
```

5 Résultats expérimentaux et Discussion

Dans cette section, nous évaluons notre approche sur un jeu de données du monde réel, *i.e.,* MOVIELENS en calculant la précision de nos recommandations pour différentes valeurs de degré de proximité afin de mettre en valeur l'utilité d'avoir des informations supplémentaires sur les utilisateurs durant le processus de recommandation ((Baeza-Yates & Ribeiro-Neto, 1999)) ((Herlocker *et al.*, 2004)). De plus, nous comparons les différentes précisions obtenues avec notre approche avec les travaux pionniers qui ont un objectif commun avec la nôtre, *i.e.,* ceux de Bellogin *et al.* (Bellogín *et al.* (2013)) et Qumsiyeh *et al.* (Qumsiyeh & Ng (2012)). Ces approches n'utilisent pas de quatrième dimension mais font appel au profil des utilisateurs comme information complémentaire pour la tâche de recommandation.

Le jeu de données filmographique MOVIELENS (*http ://movielens.umn.edu/*) est un système de recommandation et un site web communautaire qui permet aux utilisateurs de partager des films en les annotant par des tags. Le jeu de données, utilisé pour nos expérimentations, est téléchargeable gratuitement (*http ://www.grouplens.org/node/73*) et contient 95580 tags appliqués à 10681 films par 71567 utilisateurs (par exemple, <*Alex*, *X-Files*, *sciencefiction*>). Le choix du jeu de données MOVIELENS est expliqué par le fait qu'en plus d'être très utilisé dans le domaine de recommandation, ce jeu de données offre des informations supplémentaires sur les utilisateurs : l'âge, la profession ou le genre.

Utilisateur	Tag	Ressource	Profil
Mulder	*action*	*X-Files*	*student*
Mulder	*sciencefiction*	*X-Files*	25 *years old*
Scully	*adventure*	*Jurassic Park*	*professor*
Scully	*bestmovie*	*Jurassic Park*	*female*
Skinner	*thriller*	*Carrie*	*Canada*
⋮	⋮	⋮	⋮

TABLE 2 – Un instantané du jeu de données MOVIELENS.

Afin d'étudier l'influence des informations supplémentaires sur les utilisateurs durant la recommandation, nous avons choisi, dans ce qui suit, le profil des utilisateurs pour modéliser la variable v dans la *v-folksonomie*. Ainsi, nous considérons désormais le degré de proximité entre deux utilisateurs comme le nombre d'informations de profil qu'ils ont en commun (par exemple, le même âge et la même profession, si le degré de proximité est égal à 2). À cet effet, les informations supplémentaires sur les utilisateurs qui sont disponibles dans MOVIELENS sont le **genre** de l'utilisateur (masculin ou féminin), sa **profession** (au nombre de 21, qui peut être éducateur, écrivain, étudiant, scientifique, etc.) ou encore l'**âge** des utilisateurs qui est divisé en cinq tranches : (*i*) $7-18$ ans ; (*ii*) $19-24$ ans ; (*iii*) $25-35$ ans ; (*iv*) $36-45$ ans et (*v*) $46-73$ ans.

Base d'apprentissage/Base de Test

Pour nos expérimentations, nous avons utilisé le protocole de validation "5-validation croisée" ((Weiss & Kulikowski, 1991)) afin d'évaluer la pertinence de notre approche. Le jeu de

données MOVIELENS a été partitionné en deux échantillons : un échantillon aléatoire contenant 80% des utilisateurs a été utilisé comme **base d'apprentissage** et un échantillon aléatoire contenant les 20% d'utilisateurs restants, a été utilisé pour la validation de nos tests (*i.e.,* **base de test**). Pour chaque utilisateur du deuxième échantillon (*i.e.,* utilisateur test), 20% aléatoires de ses tags et ressources sont considérées comme ensemble de test/réponse et 80% comme son ensemble d'apprentissage. Nous avons répété cette expérience cinq fois en changeant à chaque fois les 20% représentant la base de test afin de couvrir les 100% de tout l'ensemble. Pour chaque utilisateur test, notre algorithme de recommandation génére une liste d'éléments (utilisateurs, tags ou ressources) en se basant sur son ensemble d'apprentissage. Si un élément de la liste de recommandation se trouve également dans l'ensemble de test de cet utilisateur, alors l'élément est considéré comme **pertinent**. Pour nos expérimentations, nous avons également fait varier le nombre de recommandations fournies à l'utilisateur : il s'agit des top-k recommandations. Grâce à ça, l'utilisateur peut spécifier les k recommandations les plus pertinentes que le système doit lui retourner. Les k premières réponses sont ceux qui ont les scores les plus élevés (*cf.*, Équation 1).

Score de ranking

Dans le but d'améliorer la précision et le rappel des recommandations proposées dans la littérature, nous proposons un nouveau score de ranking afin de classer les différentes recommandations. Pour un jeu de données donné, les top-k recommandations consistent en une liste d'items classés par valeur de score décroissante. Dans ce qui suit, la fonction de score est définie pour la recommandation de ressource mais peut très bien être définie pour la recommandation de tags ou d'utilisateurs en changeant les variables de l'équation. Ainsi, pour générer une recommandation de ressource pour un utilisateur donné, nous calculons le ranking comme décrit ci-dessus, et nous restreignons les résultats aux top-k premiers résultats (avec les scores les plus élevés). La mesure de score (notée rec_score) correspondant à un ensemble d'informations de profil V est défini comme suit :

$$\textit{rec\_score}(r_i, V) = \frac{|u_i|}{|UU|} / \exists t_i \exists r_i \exists v_i \in V, (u_i, t_i, r_i, v_i) \in \mathcal{F}_v \tag{1}$$

Donc, le score rec_score d'une ressource r_i correspondant à un profil v est le nombre d'utilisateurs uniques, ayant le même profil v (ou au moins une information de profil $v_i \in v$), qui ont partagé cette ressource, divisé par le nombre total d'utilisateurs uniques dans l'ensemble des quadri-concepts fréquents (noté UU). Par exemple, si une ressource r_1 a été partagée par 7 différents utilisateurs (au même profil) parmi une liste de 67 utilisateurs uniques, son score sera égal à 0.104 alors qu'une autre ressource r_2 partagée par 16 différents utilisateurs (au même profil) parmi la même liste aura un score égal à 0.238.

Évaluation des recommandations

Le Tableau 3 montre les valeurs de précision des recommandations obtenues par notre système de recommandation pour différents degrés de proximité et différentes valeurs de k [3] allant

3. le nombre de recommandations retournées à l'utilisateur.

Information de profil / k	6	7	8	9	10	**Précision Moyenne**	**Variance**	**Écart Type**
Degré de proximité = 0								
Aucun	0.56	0.54	0.51	0.48	0.48	0.514	0.000922	0.030364
Degré de proximité = 1								
Âge	0.60	0.57	0.54	0.51	0.50	0.544	0.001447	0.038042
Localisation	0.72	0.73	0.75	0.74	0.71	0.730	0.000200	0.014142
Profession	0.55	0.50	0.51	0.50	0.50	0.512	0.000260	0.016149
Degré de proximité = 2								
Âge + Localisation	0.52	0.52	0.52	0.51	0.51	0.516	0.000019	0.004358
Profession + Localisation	0.53	0.51	0.50	0.44	0.42	0.480	0.001800	0.042426
Âge + Profession	0.63	0.64	0.63	0.64	0.67	0.642	0.000241	0.015543
Degré de proximité = 3								
Âge + Profession + Localisation	0.50	0.42	0.37	0.33	0.30	0.384	0.004938	0.070270
Approches de la littérature								
Bellogin *et al.*	0.40	0.37	0.35	0.33	0.32	0.354	0.000824	0.028705
Qumsiyeh *et al.*	0.27	0.27	0.25	0.24	0.23	0.252	0.000256	0.016000

TABLE 3 – Valeurs de précision des recommandations pour différents degrés de proximité pour le jeu de données MOVIELENS (*cf.,* Figure 1).

de 6 à 10 sur le jeu de données MOVIELENS. Tout d'abord, les résultats démontrent l'utilité de la quatrième dimension, *i.e.,* le profil, durant le processus de recommandation. En effet, les autres meilleurs scores de précision sont atteints lorsque notre système de recommandation prend le profil des utilisateurs comme information supplémentaire. Ainsi, le recours aux informations supplémentaires sur les utilisateurs permet de personnaliser les recommandations et de générer des recommandations plus ciblées. De plus, pour toutes les valeurs de k, notre système de recommandation obtient une meilleure précision que celles de Bellogin *et al.* et Qumsiyeh *et al.* La différence est encore plus grande lorsque nous prenons en compte des informations supplémentaires sur les utilisateurs. Ainsi, RECOMMENDER améliore la précision des approches de Bellogin *et al.* et Qumsiyeh *et al.* de, respectivement, 48.49% et 140.48%. L'information de profil la plus influente est la **localisation** des utilisateurs. En effet, plus les utilisateurs sont proches géographiquement, plus ils ont tendance à partager les mêmes ressources et à avoir le même comportement social selon les traditions culturelles de leurs pays respectifs. Par exemple, les utilisateurs indiens partagent les films de Bollywood tandis que les utilisateurs japonais ont tendance à partager en masse les mangas. Ensuite, la seconde information de profil la plus influente pour la recommandation est l'**âge** des utilisateurs. En effet, les utilisateurs appartenant à la même catégorie d'âge convergent vers un vocabulaire commun (les jeunes utilisateurs contre les anciens utilisateurs) et partagent le même type de ressources, par exemple, les jeunes utilisateurs préfèrent les films d'actions, les plus jeunes partagent les mangas alors que les plus anciens ont tendance à partager les films classiques. Par ailleurs, lorsque nous associons deux informations de profil, l'âge apparaît comme étant l'information de profil la plus importante,

notamment lorsqu'elle est associée à la profession ou la localisation. En effet, la précision de notre système de recommandation augmente sensiblement lorsque nous prenons en compte à la fois l'âge et la profession comme informations supplémentaires étant donné que les utilisateurs d'une même catégorie d'âge et exerçant le même métier ont un profil assez proche, par exemple, des étudiants de [19-24] ans ou encore des techniciens de [25-35] ans.

Toutefois, la précision de nos recommandations atteint ses moins bons résultats lorsque nous combinons toutes les informations de profil. Si la localisation ou la combinaison âge-profession produit de bons résultats en termes de précision, les combiner réduit considérablement la qualité des recommandation. En effet, le nombre de recommandations décroît étant donné qu'il est rare de trouver des utilisateurs ayant à la fois la même profession, le même âge, la même localisation et partageant les mêmes ressources. Comme il es rare de retrouver des utilisateurs ayant ce même genre de profil, les ressources recommandées ont moins de chance d'être pertinentes. Par ailleurs, si l'âge ou la localisation donnent de bons résultats en termes de précision, cela n'est pas le cas pour la profession qui n'est pas une information influente pour nos recommandations. Les utilisateurs exerçant le même métier ne partagent pas forcément les mêmes intérêts. Enfin, lorsque notre système de recommandation ne prend aucune information supplémentaire sur les utilisateurs, la précision décroît rapidement puisque la liste de recommandation est aléatoire, c'est-à-dire, pas personnalisée. Ainsi, dans le cas des *v-folksonomies*, plus le nombre de ressources recommandées augmente, moins elles sont pertinentes. En effet, les ressources les plus partagés par le passé ne sont pas nécessairement partagés dans le futur, donc, le score de précision n'est pas élevé lorsqu'aucune information supplémentaire sur les utilisateurs n'est prise en compte. Nous concluons que prendre en compte des informations supplémentaires sur les utilisateurs permet d'augmenter la précision des recommandations, et pour avoir les meilleurs résultats, il est préférable de s'arrêter à une ou deux informations de profil. Ainsi, si nous prenons une seule information de profil, la localisation et l'âge sont les informations de profil qui donnent les meilleurs scores. Par contre, si nous combinons deux différentes variables, il est conseillé de combiner l'âge avec une autre information de profil.

6 Conclusion et Perspectives

Dans ce papier, nous avons considéré une nouvelle dimension dans une *folksonomie* contenant des informations supplémentaires sur les utilisateurs. Ensuite, nous avons proposé notre système personnalisé de recommandation qui repose sur une mesure de proximité entre utilisateurs afin d'améliorer la qualité des recommandations. Les expérimentations ont montré l'utilité d'avoir des informations supplémentaires durant le processus de recommandation. Parmi nos perspectives de recherche, nous cherchons à étendre les informations supplémentaires aux reviews, commentaires et l'historique de recherche des utilisateurs afin d'avoir un suivi dynamique des utilisateurs et améliorer encore plus les recommandations.

Remerciements.

Ce travail est partiellement financé par le projet franco-tunisien PHC Utique 11G141. Nous remercions les relecteurs anonymes pour leurs remarques constructives.

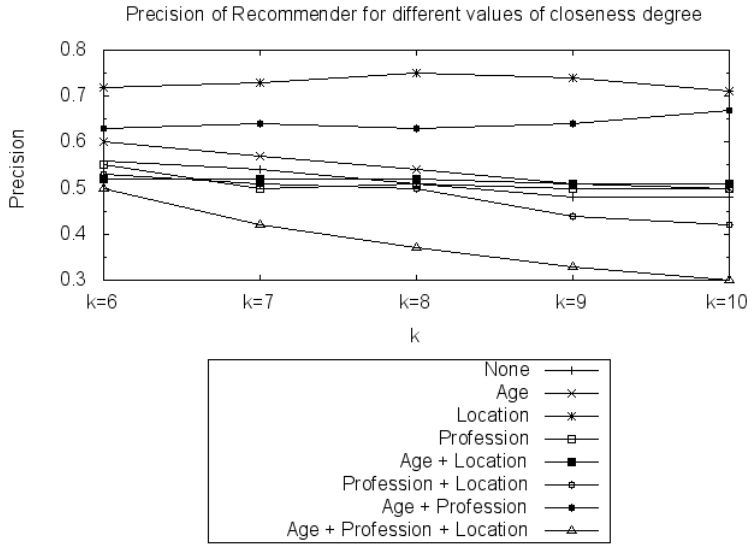


FIGURE 1 – Valeurs de précision des recommandations pour différents degrés de proximité pour le jeu de données MOVIELENS.

Références

BAEZA-YATES R. A. & RIBEIRO-NETO B. (1999). *Modern Information Retrieval*. Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc.

BASILE P., GENDARMI D., LANUBILE F. & SEMERARO G. (2007). Recommending smart tags in a social bookmarking system. In *Bridging the Gap between Semantic Web and Web 2.0*, p. 22–29.

BELLOGÍN A., CANTADOR I. & CASTELLS P. (2013). A comparative study of heterogeneous item recommendations in social systems. *Inf. Sci.*, **221**, 142–169.

CERF L., BESSON J., ROBARDET C. & BOULICAUT J.-F. (2009). Closed patterns meet n-ary relations. *ACM TKDD*, **3**, 3 :1–3 :36.

DE MEO P., QUATTRONE G. & URSINO D. (2010). A query expansion and user profile enrichment approach to improve the performance of recommender systems operating on a folksonomy. *User Modeling and User-Adapted Interaction*, **20**(1), 41–86.

DIEDERICH J. & IOFCIU T. (2006). Finding communities of practice from user profiles based on folksonomies. In *Proceedings of the 1st International Workshop on TEL-CoPs, Crete, Greece*, p. 288–297.

HERLOCKER J. L., KONSTAN J. A., TERVEEN L. G. & RIEDL J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, p. 5–53.

HOTHO A., JÄSCHKE R., SCHMITZ C. & STUMME G. (2006). Information retrieval in folksonomies : Search and ranking. In *Proc. of ESWC, Budva, Montenegro*, volume 4011 of *LNCS*, p. 411–426 : Springer, Heidelberg.

HU J., WANG B. & TAO Z. (2011). Personalized tag recommendation using social contacts. In *Proc. of Workshop SRS'11, in conjunction with CSCW*.

JÄSCHKE R., HOTHO A., SCHMITZ C., GANTER B. & STUMME G. (2008). Discovering shared conceptualizations in folksonomies. *Web Semantics.*, **6**, 38–53.

JÄSCHKE R., MARINHO L., A. HOTHO A., LARS S.-T. & STUMME G. (2007). Tag recommendations in folksonomies. In *Proc. of the 11th ECML PKDD, Warsaw, Poland*, p. 506–514.

JELASSI M. N., BEN YAHIA S. & MEPHU NGUIFO E. (2013). A personalized recommender system based on users' information in folksonomies. In *Proc. of the 22nd International Conference on World Wide Web companion*, WWW '13 Companion, p. 1215–1224.

JELASSI M. N., BEN YAHIA S. & MEPHU NGUIFO E. (2015). Towards more targeted recommendations in folksonomies. *Social Netw. Analys. Mining*, **5**(1), 68 :1–68 :18.

LANDIA N. & ANAND S. (2009). Personalised tag recommendation. *Recommender Systems & the Social Web, New York, NY, USA*, p. 83–86.

LIANG H., XU Y., LI Y. & NAYAK R. (2010). Personalized recommender system based on item taxonomy and folksonomy. In *Proceedings of the 19th ACM CIKM'10*, p. 1641–1644, New York, NY, USA : ACM.

LIPCZAK M. (2008). Tag recommendation for folksonomies oriented towards individual users. In *Proc. of the ECML/PKDD Discovery Challenge, Antwerp, Belgium*, p. 84–95.

MIKA P. (2007). Ontologies are us : A unified model of social networks and semantics. *Journal of Web Semantics.*, **5**(1), 5–15.

NANOPOULOS A., RAFAILIDIS D., SYMEONIDIS P. & MANOLOPOULOS Y. (2010). Musicbox : Personalized music recommendation based on cubic analysis of social tags. *Trans. Audio, Speech and Lang. Proc.*, **18**(2), 407–412.

QUMSIYEH R. & NG Y.-K. (2012). Predicting the ratings of multimedia items for making personalized recommendations. In *SIGIR'12*, p. 475–484, New York, NY, USA : ACM.

F. RICCI, L. ROKACH, B. SHAPIRA & P. B. KANTOR, Eds. (2011). *Recommender Systems Handbook*. Springer.

STROHMAIER M., KÖRNER C. & KERN R. (2012). Understanding why users tag : A survey of tagging motivation literature and results from an empirical study. *Web Semant.*, **17**, 1–11.

WEISS S. M. & KULIKOWSKI C. A. (1991). *Computer Systems That Learn : Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

Améliorer la Recommandation de Ressources dans les Folksonomies par l'Utilisation de Linked Open Data

Samia Beldjoudi[1,2], Hassina Seridi[2], Abdallah Benzine[2]
[1] Ecole Préparatoire Aux Sciences et Techniques Annaba, Algérie
[2] Laboratorie de Géstion Electronique de Documents LabGED, Université Badji Mokhtar Annaba, Algérie
{beldjoudi, seridi}@labged.net
abdoulahbenzine@gmail.com

Abstract : Le Web social permet aux utilisateurs de créer, annoter, partager et rendre public les ressources qu'ils jugent intéressantes. Les folksonomies tiennent une place importante dans ces nouvelles pratiques sociales et sont utilisées dans de nombreuses applications dont les systèmes de recommandation. Aussi, l'émergence des Linked Open Data (LOD) permet d'établir des liens entre différentes entités issues de diverses sources en connectant les informations dans un unique espace de données. Dans ce travail, en plus de la prise en compte des interactions sociales afin de surmonter les problèmes d'ambigüité des tags, nous montrons comment le contenu structuré disponible à travers les LOD peut être utilisé. Les LOD sont en effet exploitées afin de pallier au manque de caractéristiques sur les ressources dans les folksonomies et faire des recommandations pertinentes et diversifiées.

Mots Clés : Folksonomies, Recommandation, Ambiguïté, Linked Open Data, Démarrage à froid, Diversité

1 Introduction

Les systèmes d'étiquetages sociaux ont gagné en popularité ces dernières années sur le Web au vu de leur simplicité pour catégoriser et retrouver les contenus en utilisant les tags. Le nombre croissant d'utilisateurs fournissant des informations sur eux-mêmes à travers leurs activités d'étiquetage sociales a induit l'émergence d'approches de profilage fondées sur les tags, qui supposent que les utilisateurs exposent leurs préférences sur des contenus au travers de tags. De ce fait, les tags peuvent être utilisés pour construire des recommandations. D'un autre côté, l'objectif principal de chaque système de recommandation est de satisfaire les intérêts des utilisateurs. L'approche classique pour cette tâche est de prédire des scores pour les ressources qui n'ont pas été évaluées par les utilisateurs et les présenter suivant l'ordre décroissant de leurs scores. Par ailleurs, ce mécanisme seul n'est généralement pas suffisant pour satisfaire les intérêts des utilisateurs. Par exemple, si un système recommande les ressources en fonction de leur popularité, il ne représente pas une plus-value significative pour l'utilisateur. En effet, les ressources ainsi recommandées à l'utilisateur ont de fortes chances d'être déjà connues par l'utilisateur puisqu'il y a de fortes chances qu'il en ait déjà entendu parler. Les critères de nouveauté et de diversité doivent être également pris en compte dans l'évaluation de la qualité d'un système de recommandation et la précision seule ne donne qu'un aperçu très partiel de l'utilité des systèmes réels.
Par ailleurs, le challenge à relever dans les systèmes de recommandation reste le problème de démarrage à froid pour les nouvelles ressources qui n'ont aucune évaluation ou pour les nouveaux utilisateurs pour lesquels le système n'a pas suffisamment d'informations.
D'autre part, les données liées (*linked data*) désignent des données suivants un paradigme fondé sur quatre règles simples : les URIs pour identifier les entités, les URLs permettant le référencement des entités, fournissant des informations utiles à ces URI fondées sur des

formats standard et la connexion et l'interconnexion à d'autres entités afin de permettre une exploration plus approfondie (Berners- Lee 2007).
Pour les données, pour être qualifiées pleinement de 'Open Linked Data' (LOD), elles devront en outre être fournies au public, disponibles sur le Web et être sous une licence ouverte. Faisant usage de formats du Web sémantique, les (LOD) mettent en œuvre la vision d'un réseau de données. Les technologies sous-jacentes permettent d'une part l'identification unique des entités via les URIs ainsi qu'une sémantique claire des relations modélisées par les liens entre les entités. Les (LOD) ont connu une croissance phénoménale au cours des dernières années. Le graphe distribué résultant des entités liées entre elles sur le web est communément appelé le nuage des (LOD) et couvre des centaines de sources de données fournissant des milliards de triplets RDF.
Ainsi, les (LOD) couvrent différents domaines allant des contenus liés aux médias, les réseaux sociaux et les contenus générés par les utilisateurs, les données bibliographiques, les sciences de la vie, la médecine, la biologie, les données géographiques, les données gouvernementales. En outre, certaines sources de données telles que DBPedia fournissent des informations générales, inter-domaines et ainsi jouent un rôle clé dans la connexion des données provenant des domaines très différents.
Dans ce papier, nous considérons le domaine du Web social sémantique et particulièrement les problèmes liés à la recommandation de ressources dans les folksonomies. Nous proposons une méthode pour analyser les profiles des utilisateurs selon leurs activités d'étiquetage afin d'améliorer la recommandation des ressources. L'efficacité de la recommandation dépend de la résolution des problèmes inhérents aux folksonmies. Dans notre processus de recommandation, nous montrons comment le problème de l'ambigüité peut être réduit en tenant compte des similarités sociales calculées sur les folksonomies combinées avec les similarités entre ressources dans les LOD. Nous utilisons également la force des LOD pour diversifier la recommandation dans les systèmes d'étiquetages sociaux et ce par l'exploration d'entités inter-liés.
Ce papier est organisé comme suit : la section 2 est un survol des travaux connexes. La section 3 est dédiée à la présentation de l'approche. Dans la section 4, les résultats sur les expérimentations sont présentées et discutées pour conclure sur les performances de notre approche. Les conclusions et les perspectives sont décrites dans la section 5.

2 Travaux connexes

Beaucoup de recherches dans le passé ont proposé d'utiliser les ontologies et les taxonomies pour améliorer la qualité des systèmes de recommandation conventionnels (Maidel et al, 2008, Middleton et al., 2004, Anand et al., 2007). Ces dernières années, avec l'émergence des LOD, une nouvelle classe de systèmes de recommandation a vu le jour nommée systèmes de recommandation basés sur les LOD. La communauté du Web sémantique et des systèmes de recommandation s'intéressent de plus en plus à cette nouvelle topologie de systèmes de recommandation. La plus part des travaux liés à ces thématiques ont essayé de réutiliser et d'adapter quelques idées issues des systèmes de recommandation ontologiques aux LOD en s'adaptant à leurs caractéristiques propres alors que d'autres ont proposé de nouvelles approches conçues spécifiquement pour les technologies des Linked Open Data et ont alors proposé de nouvelles applications des systèmes de recommandation pour celles-ci. Dans ce qui suit, nous allons passer en revue les contributions les plus significatives. L'une des approches qui exploite les Linked Open Data pour construire des systèmes de recommandation est celle de (Marie et al, 2013) dans laquelle des datasets des LOD sont utilisés pour une exploration personnalisée utilisant une méthode d'activation en diffusion. Une méthode d'activation en diffusion a été utilisée afin de trouver des relations sémantiques

entre des items appartenant à différents domaines. Un système de recommandation entièrement basé sur SPARQL nommé RecSPARQL a été présenté dans (Ayala et al, 2014). L'outil proposé étend la syntaxe et la sémantique de SPARQL afin de permettre un filtrage collaboratif flexible et générique et une recommandation basée contenu sur des graphes RDF. Dans (Khrouf et Troncy, 2013), les auteurs présentent un système de recommandation événementiel basé sur les Linked Data et la diversité des utilisateurs. Une extension sémantique du modèle SVD+++ nommé SemanticSVD+++ est présentée dans (Rowe, 2014). Elle intègre des catégories sémantiques d'items dans le modèle. Ce modèle est également capable de considérer l'évolution au fil du temps des préférences des utilisateurs. Dans (Rowe, 2014), les auteurs améliorent le travail précédent pour tenir compte des items démarrant à froid. Ils introduisent des sommets-noyaux afin d'obtenir des informations sur les catégories sémantiques non évaluées en démarrant des catégories connues. Enfin, dans (Dojchinovski et Vitvar, 2014) les auteurs proposent l'utilisation de techniques de recommandation afin de fournir un accès personnalisé aux Linked Data. La méthode de recommandation proposée est un système de filtrage collaboratif utilisateur-utilisateur où la similarité entre les utilisateurs prend en compte les points communs et l'informativité des ressources au lieu de les considérer comme de simples identificateurs.

D'autre part, dans les systèmes d'étiquetage social, l'objectif général de la recommandation de ressources est d'assurer la quantité et la pertinence des ressources recommandées. Parmi les travaux traitant de ce problème, nous pouvons citer (Huang et al, 2011) qui a proposé un système de recommandation qui utilise les préférences les plus récemment identifiées dans les tags des utilisateurs. (Zanardi et al, 2011) ont proposé une méthode destinée à étendre les capacités de recherche des collections digitales ciblant les domaines académiques et scolaires. (Beldjoudi et al., 2011, 2012) ont proposé une méthode pour analyser les profiles des utilisateurs afin d'améliorer la recommandation de ressources dans les folksonomies. L'objectif est d'enrichir les profils des utilisateurs avec des ressources pertinentes en résolvant le problème de l'ambigüité des tags durant la recommandation.

Le problème de la diversité des résultats a déjà été traité dans la Recherche d'Information(RI) mais sous un angle diffèrent. Ce problème est traité par la (RI) afin de résoudre celui de l'ambigüité et/ou de la sous-spécification des requêtes des utilisateurs. Dans la Recherche d'Information, l'accent est mis sur l'élargissement des items recommandés présentés à l'utilisateur (diversité) et la promotion d'items moins connus (nouveauté) ou d'items non familiers pour un utilisateur donné. Quelques recherches ont été menées dans ce terrain, et ont connu un intérêt croissant à cause de l'importance de la diversité et de la nouveauté dans la communauté des systèmes de recommandation. Néanmoins, il reste encore un espace de recherche considérable dans l'amélioration de la recommandation de ressources dans les systèmes d'étiquetage social par l'utilisation des Linked Open Data afin d'assurer des résultats précis et diversifiés.

3 Description de l'approche

Une folksonomie est définie comme un modèle triparti où les ressources Web sont associées à un utilisateur par une liste de tags. Formellement, une folksonomie est un tuple F = < U, T, R, A> où U, T et R représentent respectivement un ensemble d'utilisateurs, un ensemble de tags et un ensemble de ressources et A représente les relations entre les trois éléments précédents, c'est-à-dire $A \subseteq U \times T \times R$ (Mika, 2005)..

Nous extrayons trois réseaux sociaux à partir d'une folksonomie, qui représentent trois points de vue différents sur les interactions sociales: un réseau relatif aux tags et aux utilisateurs et un second concernant les tags et les ressources et un troisième concernant les utilisateurs et les ressources. Nous représentons ces réseaux sociaux par trois matrices TU, TR, UR :

-TU = [X_{ij}] où : $X_{ij}= \begin{cases} 1 \text{ si } \exists \text{ r} \in \text{R}, < uj, ti, r > \in \text{A} \\ 0 \text{ autrement} \end{cases}$

-TR = [Y_{ij}] où: $Y_{ij} = \begin{cases} 1 \text{ si } \exists \text{ u} \in \text{U}, < u, ti, rj > \in \text{A} \\ 0 \text{ autrement} \end{cases}$

-UR = [Z_{ij}] où: $Z_{ij} = \begin{cases} 1 \text{ si } \exists \text{ t} \in \text{T}, < ui, t, rj > \in \text{A} \\ 0 \text{ autrement} \end{cases}$, RU, RT et UT sont transposées dans les matrices UR, TR and TU.

C'est ce qui nous permet d'analyser les corrélations issues des différentes interactions sociales. Nous utilisons Pajek, un outil qui a déjà été utilisé par Mika pour analyser les grands réseaux (Mika, 2005).

Dans cet article, nous proposons une méthode pour analyser les profils des utilisateurs d'après leurs tags afin de trouver des ressources intéressantes et les recommander. L'objectif est d'enrichir les profils des utilisateurs de folksonomies avec des ressources pertinentes. Nous supposons que le partage automatique de ressources renforce les liens sociaux entre les acteurs et nous exploitons cette idée afin de réduire l'ambigüité des tags dans le processus de recommandation en augmentant le poids associé aux ressources web selon les similarités sociales. Nous nous sommes basés sur des règles d'association qui sont une méthode puissante pour découvrir des corrélations intéressantes entre un grand ensemble de données sur le Web. Pour appliquer une méthode de règle d'association dans les folksonomies, nous avons représenté chaque utilisateur dans la folksonomie par un ID de transaction et les tags qu'ils utilisent par l'ensemble des éléments qui sont dans cette transaction (Beldjoudi et al., 2012).

Notre objectif est de trouver des corrélations entre les balises, c.à.d. de trouver des tags apparaissant fréquemment ensembles afin d'en extraire ceux qui ne sont pas utilisés par un utilisateur particulier, mais qui sont souvent utilisés par d'autres utilisateurs proches de lui. Par exemple, considérons un ensemble de données dans lequel il s'apparait que de nombreux utilisateurs utilisant le tag *Software* utilisent également le tag *Java*. Nous cherchons à extraire une règle *Software* → *Java* afin que nous puissions enrichir les profils des utilisateurs qui emploient le tag *Software*, mais pas le tag *Java*, par les ressources taggués avec *Java*. Une fois que les règles sont extraites, notre système de recommandation se déroule comme suit: Pour chaque règle extraite, nous testons si les balises qui sont dans l'antécédent de la règle sont utilisées par l'utilisateur actuel. Si tel est le cas, alors les ressources taggués avec chaque balise trouvée dans le conséquent de la règle sont candidates à être recommandée par le système.

L'efficacité de la recommandation dépend de la résolution des problèmes inhérents aux folksonomies. Dans notre approche, nous abordons les problèmes d'ambiguïté des tags, les variations orthographiques (ou synonymie) et le manque de liens sémantiques entre tags. Le détail sera décrit dans les prochains paragraphes.

3.1 Exploiter les similarités sociales et le LOD pour surmonter l'ambiguïté des tags et le démarrage à froid lors de la recommandation

Selon (Mathes, 2004), «Les problèmes inhérents à un vocabulaire non contrôlé conduit à un certain nombre de limites et les faiblesses dans les folksonomies. L'ambiguïté des tags peut être levée lorsque les utilisateurs appliquent le même tag de différentes manières".

Une balise peut avoir plusieurs significations, c.à.d. référer à plusieurs concepts. Par conséquent, un système de recommandation basé sur les tags recommande aussi bien des ressources relatives aux fruits ou aux ordinateurs à un utilisateur qui recherche avec le tag "apple". La résolution de problème d'ambiguïté est particulièrement cruciale dans notre approche, où certaines balises qui sont utilisées pour recommander des ressources ne sont pas utilisées directement par l'utilisateur mais déduit des règles d'association (Beldjoudi et al, 2011). Pour résoudre le problème d'ambiguïté lors de la recommandation, nous proposons de

mesurer la similarité entre utilisateurs afin d'identifier ceux qui ont des préférences similaires et par conséquent adapter la recommandation aux profils d'utilisateurs (voir Algorithme 1).
Nous expliquons comment les similarités sociales et les LOD sont utilisés pour surmonter l'ambiguïté des tags et le problème de démarrage à froid lors de la recommandation dans ces deux étapes :
- *Première étape*: Pour chaque règle d'association A → B dont l'antécédent s'applique à un utilisateur actif u_x , nous mesurons les similarités entre cet utilisateur et les utilisateurs qui utilisent les tags qui se trouvent dans le conséquent de la règle. Les ressources associées à ces tags sont recommandées à cet utilisateur en fonction de ces similarités. Pour mesurer la similarité entre deux utilisateurs u1 et u2, les deux sont représentés par un vecteur binaire représentant tout leurs tags (extrait de la matrice UT) et on calcule le cosinus de l'angle entre les deux vecteurs: $\text{sim}(u_1, u_2) = cos(v_1, v_2) = \frac{v_1.v_2}{\|v_1\|^2\|v_2\|^2}$ (1)

Selon (Cattuto et al., 2008) et (Koerner et al., 2010), le calcul de similarité avec la formule cosinus donne de bons résultats en un coût de calcul très raisonnable, car il a une complexité linéaire. Nous insistons sur le fait que la distribution des tags en fonction des ressources et des utilisateurs dans les folksonomies suit une loi de calcul de puissance: la plupart des ressources sont marquées par un petit nombre d'utilisateurs, et de nombreux tags ne sont utilisés que par quelques utilisateurs, une propriété qui conduit à une faible valeur de r (le nombre de ressources dans la matrice RU) et n (le nombre d'utilisateurs dans la matrice UT). Par conséquent, notre approche peut évoluer dans les très grandes bases de données.
- *Deuxième étape:* Pour éviter le problème de démarrage à froid qui résulte généralement du manque de données requises par le système afin de faire une bonne recommandation, lorsque l'utilisateur du système de recommandation n'est pas encore semblable à d'autres utilisateurs, nous proposons d'exploiter les liens sémantiques entre les ressources dans le LOD. Celles-ci peuvent être considérées comme une source fiable et riche d'informations. Elles aident les systèmes de recommandation à résoudre certains problèmes, tels que le problème du démarrage à froid et l'analyse de contenu limité. On se fonde pour cela sur une mesure robuste des similarités entre les ressources en utilisant les LOD. Dans cette approche, nous utilisons le Open Linked Data pour apprécier la similarité entre les ressources d'une folksonomie en utilisant leurs ressources correspondantes sur les LOD (Fig1) (c.à.d. nous mesurons la similarité entre les ressources qui seraient recommandées par le système (celles qui sont liées à un tag apparaissant dans la conséquence d'une règle d'association) et celles qui sont déjà recommandées à l'utilisateur. La similarité entre deux ressources est calculée en utilisant l'indice de Jaccard défini comme suit: $sim(R1, R2) = J(R1, R2) = \frac{|R1 \cap R2|}{|R1 \cup R2|}$ (2)

Chaque ressource Rx est définie par ses caractéristiques c.à.d. des triplets de type (Rx, prédicat, Ry), où prédicat indique le type de la relation et Ry représente le nœud cible (c.à.d. le nœud connecté à l'autre extrémité de la relation).

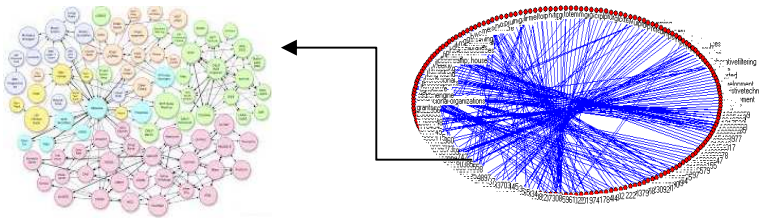


FIGURE 1 -Lier les ressources dans la base del.icio.us (qui sont représentées par l'outil Pajek) à leurs ressources correspondantes dans DBpedia

Algorithme1 : Recommandation personnalisée de ressources

Entrée: Une folksonomie : F< U, T, R, A>, *S1, S2*: des nombres entiers positifs
Sortie: L_r: liste de ressources à recommandées

Début

1. Génération de N Règles associative $\{t_A \rightarrow t_B\}$

2. Pour K=1 jusqu'à N faire,

3. Construire la matrice $UR|t_B = [Z_{ij}]$ où: $Z_{ij} = \begin{cases} 1 \text{ if } < u_i, t_B, r_j > \in A \\ 0 \text{ otherwise} \end{cases}$

4. Construire la matrice $UT|r_j = [X_{ij}]$ où : $X_{ij} = \begin{cases} 1 \text{ if } < u_x, ti, r_j > \in A \\ 0 \text{ otherwise} \end{cases}$

5. Construire la matrice UU= UT*TU

6. Calculer Sim-u= Cos (v1,v2)

7. Si Sim-u >= S1 alors, $L_r = L_r \cup \{rj\}$

8. Sinon Calculer Sim-r = $J(rj, rm)$ en utilisant LOD

9. Si Sim-r >= S2 L=L ∪ {rj}

10. Fin Si

11. Fin Sinon

12. Fin Si

13. Fin Pour

14. Renvoyer (L_r);

Fin

3.2 Assurer la diversité dans la Recommandation

Lors de l'utilisation d'un système de recommandation tels que Amazon.com, Netflix, etc. on peut rencontrer le problème suivant: si un profil d'utilisateur est composé d'un couple de livres de "Victor Hugo", un moteur de recommandation axée uniquement sur la précision peut fournir une liste composée principalement d'autres livres de "Victor Hugo". Bien qu'il soit très probable que l'utilisateur va aimer les livres recommandés, il est clair que la recommandation n'est pas très utile dans le sens de:

-Le Manque de diversité, probablement un plus petit échantillon de livres de "Victor Hugo" aurait été aussi utile pour découvrir le travail de l'auteur et aurait donné l'espace pour d'autres livres intéressants pour d'autres auteurs; et

-Le Manque de nouveauté, puisque "Victor Hugo" est un auteur très connu, pour lequel un système de recommandation n'est même pas nécessaire.

Cette situation ouvre deux questions: Pourquoi le système fournit de tels résultats ? Comment résoudre ce problème ? En règle générale, les systèmes de recommandation sont entrainés à minimiser l'erreur de prédiction, de sorte que des aspects tels que la redondance et l'évidence ne sont généralement pas considérés. Un autre problème réside dans la sous-spécification du profil d'utilisateur. Comme il contient qu'un livre d'un auteur unique, une approche de filtrage collaboratif pur est susceptible de trouver la plupart des connexions à d'autres utilisateurs qui auront plus de livres du même auteur. Enfin, même si l'utilisateur avait acheté ou naviguer vers des livres d'autres auteurs, ceux de "Victor Hugo" resteront toujours populaires et donc seront inévitablement favorisés par un algorithme de recommandation standard.

Pour résoudre ce dilemme dans les folksonomies, nous proposons d'extraire à partir des ressources les caractéristiques les plus populaires trouvés dans le profil de l'utilisateur (c.à.d. les caractéristiques qui intéressent l'utilisateur au moment de choisir ses ressources) et ensuite explorer le graphe de LOD afin d'en extraire des ressources liées à ces caractéristiques. Par exemple, considérons le cas suivant:

FIGURE 2 - Le profil de l'utilisateur contenant un ensemble de films de Leonardo Dicaprio

Dans cet exemple, le profil de l'utilisateur est composé des ressources (R1, R2, R3, R4 et R5), dont l'intersection entre les caractéristiques de ces ressources doit être calculé (R1∩ R2 R3 ∩

∩R4 ∩ R5). On extrait ainsi les caractéristiques les plus populaires qui intéressent l'utilisateur quand il choisit de tagguer ses propres ressources. Ensuite, pour chaque caractéristique (Pi) dans le résultat de l'intersection, nous allons explorer le graphe LOD à trois niveaux pour extraire d'autres ressources (R6) ayant cette caractéristique ou ayant un lien direct / indirect avec ces dernières (R7, R8 resp) .Nous avons fixé le niveau d'exploration à 3 afin d'éviter de biaiser les résultats de recommandation.
Nous avons dans l'exemple ci-dessus (R1∩ R2 ∩ R3 ∩R4 ∩ R5) = {Leonardo Dicaprio ...}. En explorant le sous-graphe suivant, nous constatons que la ressource "Leonardo DiCaprio" est liée à la ressource "OSCARS" via le prédicat (has). A son tour, la ressource "OSCARS" est lié via le prédicat (winner) avec la ressource "Eddie Redmayne". Par conséquent, nous pouvons recommander des films de « Eddie Redmayne » par exemple à l'utilisateur actuel.

FIGURE 3 – Un sous graphe de LOD

Avec cette méthode, nous nous assurons que la liste des ressources à recommander est diversifiée, où chaque utilisateur peut obtenir autres ressources différentes à celles qui se trouvent dans son profile, même si elles ne figurent pas dans les profils de ses voisins dans le réseau social.
Chaque ressource recommandée par le système est d'abord associée un poids initial basé sur les similarités entre utilisateurs. Au-dessus d'un seuil fixé dans [0..1], nous qualifions la ressource comme fortement recommandée. Sous ce seuil, nous considérons la similarité entre ressources et nous recommandons fortement de même les ressources que les poids calculés sur LOD sont au-dessus d'un seuil donné. Nous notons que notre système de recommandation est flexible, puisque l'utilisateur peut interagir pour accepter ou rejeter les ressources recommandées.

3 Les résultats expérimentaux

Nous avons montré dans la section précédente que notre approche utilise les dimensions sociales et les sémantiques du web afin d'améliorer le processus de recommandation. Cette section donne des détails sur l'implémentation pour permettre son évaluation et montrer son efficacité.
Afin de valider notre approche, nous avons conduit une expérimentation avec la base del.icio.us. Notre base de test comprend 58588 assignations de tags impliquant 12780 utilisateurs, 30500 tags parmi lesquels certains sont ambigus et ont des orthographes différentes et 14390 ressources chacune étant associée à plusieurs tags et à plusieurs utilisateurs. Notre système a extrait un ensemble de 946 règles d'association de la base de données avec un support égal à 0.5 et une confiance égale à 0.6.
La principale base de données LOD utilisée pour nos expérimentation est DBPedia, l'une des initiatives du web sémantique ayant eu le plus de succès. Afin d'évaluer notre système de

recommandation basé sur les LOD, les ressources de la base del.cio.us doivent être mise en correspondance avec celles de DBpedia.
Les LOD peuvent être interrogées au travers de leurs endpoints SPARQL. Pour DBpedia, cela permet à n'importe qui d'effectuer des requêtes complexes sur n'importe quel sujet disponible dans Wikipedia. Par exemple, on peut savoir simplement quels acteurs ont joué dans le film « Le Revenant » via la reqête SPARQL :
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
SELECT ? actor WHERE {dbpedia:the_revenant dbpedia-owl:starring ?actor .}
Cet exemple permet de voir comment on peut extraire de nombres informations aussi bien sur une ressource spécifique que sur plusieurs d'entre elles. A partir de l'url associée à un item, il est possible d'extraire le sous-graphe associé en effectuant plusieurs requêtes SPARQL en utilisant une stratégie profondeur d'abord à profondeur limitée.
La sémantique des classes LOD et leurs relations sont décrites grâce à des ontologies. Par exemple, la ressource dbpedia :Leonard-dicaprio dans DBpedia est une instance de la classe dbpedia-owl :Person qui est à son tour une sous-classe de dbpedia-owl :Agent. La sémantique des propriétés est également définie dans une telle ontologie. Par exemple, la propriété dbpedia-owl :starring qui relie dbpedia :The_revenant à dbpedia :Leonardo-dicaprio a pour domaine dbpedia-owl :Workand et pour *range* dbpedia-owl :Actof qui est une sous-classe de dbpedia-owl :Person.

4.1 Le protocole expérimental

Idéalement, dans le but d'évaluer la qualité d'un système de recommandation, nous devons montrer que les ressources recommandées sont réellement acceptées par l'utilisateur. Mais pour le savoir, nous devrions interroger les utilisateurs des bases de données choisies et leur demander s'ils ont apprécié l'ensemble des ressources proposées. Puisque ceci est impossible, nous avons retiré aléatoirement certaines ressources du profil de chaque utilisateur et nous avons appliqué notre approche sur l'ensemble des données restantes afin de voir si les ressources retirées sont recommandées à leurs utilisateurs respectifs ou pas. Si elles sont recommandées, nous pouvons alors conclure que le système a correctement estimé les préférences de l'utilisateur. Afin de tester les performances de notre approche, nous avons suivi les étapes suivantes :

a) Evaluation de la capacité à dépasser le problème de l'ambigüité

Pour atteindre ce but, nous avons commencé par sélectionner un ensemble de 1154 tags ambigus de la base del.ici.us. Nous avons alors aléatoirement retiré des ensembles de ressources correspondants à ces tags ambigus. Ce processus a été répété cinq fois pour chaque tag dans le but d'effectuer une cross-validation. En d'autres termes, pour chaque tag, nous avons aléatoirement divisé l'ensemble des ressources correspondantes en cinq parties et nous avons ensuite sélectionné la partie à retirer dans chaque évaluation pour l'utiliser comme un ensemble de test. Ce processus a été répété cinq fois et à chaque nous avons choisi un ensemble de test diffèrent.

- Résultats expérimentaux : Pour évaluer la qualité de notre système de recommandation, nous avons utilisé trois métriques : rappel, précision et F1 qui est une combinaison des deux premières. 107 règles d'association ont été extraites avec un support égal à 0.5 et une confiance égale à 0.6. Ensuite, les trois métriques ont été calculées pour chaque participant. La table 1 présente les valeurs moyennes des métriques :

Table 1: Précision, rappel et F1 moyennes des recommandations

Précision	Rappel	F1
77%	83%	80%

Ces résultats montrent que, en appliquant les règles d'association extraites, les ressources associées à des tags non ambigus sont très recommandées. Cela montre également que, dans le cas des règles faisant intervenir des tags ambigus, notre système recommande à l'utilisateur des ressources qui sont proches de ses intérêts avec un haut niveau de recommandation et celles qui sont éloignées de ses intérêts avec un faible niveau de recommandation.

b) **Evaluation de la capacité à dépasser le problème de variations d'orthographes**

Pour atteindre ce second objectif, nous avons commencé par sélectionner un ensemble de tags contenant des termes avec beaucoup d'orthographes différentes. Cela donne un ensemble de 2417 tags extraits de la base del.icio.us. Ensuite, nous avons aléatoirement retiré des ressources étiquetées par ces tags afin de déterminer si le système les recommande aux bons utilisateurs. Ce processus a été répété cinq fois afin d'effectuer une validation croisée.

- Résultats expérimentaux : Basé sur nos ensembles de test, 127 règles d'association ont été extraites et cela avec un support égal à 0.5 et une confiance égale à 0.6. Ensuite, nous calculé les mêmes métriques que précédemment pour chaque utilisateur. La table 2 les valeurs moyennes obtenues pour chaque métrique :

Table 2 : Précision, rappel et F1 moyennes des recommandations

Précision	Rappel	F1
69%	80%	75%

4.2 Discussion

Nous pouvons conclure de l'analyse des résultats précédents que, dans tous les cas, la précision, le rappel et la métrique F1 de notre approche sont très prometteuses pour la base del.icio.us. Ces résultats indiquent que l'utilisation de règles d'association et des similitudes sociales combinées aux LOD permet de tenir compte du profil de l'utilisateur lors de la recommandation de ressources. En effet, ces résultats montrent que notre approche réussit à distinguer entre les tags ambigus et permet de tenir en compte les variations de l'orthographe durant la recommandation de ressources. La table 3 présente l'écart-type de la précision, du rappel et de la métrique F1 dans la base del.icio.us pour l'ambigüité des tags et le problème des orthographes multiples.

Table 3: L'écart-type des trios métriques pour l'ambigüité des tags et les orthographes multiples

	Précision	Rappel	F1
L'ambiguité des tags	5%	6%	5%
Orthographes multiples	8%	5%	4%

Dans les deux cas, ces valeurs sont très petites ce qui indique que les valeurs de ces trois mesures pour chaque utilisateur sont très proches de la moyenne. Les valeurs moyennes (présentés dans les tables 1 et 2) étant très prometteuses pour la communauté en général, les petites valeurs de l'écart-type indiquent que les métriques sont également très prometteuses pour chaque utilisateur.

4.3 Le Choix de la valeur optimale pour le support et la confiance

L'objectif de la fouille par les règles d'association est de trouver toutes les règles qui satisfont un support minimum et des restrictions de confiance. Plus on augmente la valeur du support, plus les règles extraites sont évidentes et alors moins elles sont utiles pour l'utilisateur. Il en résulte qu'il est nécessaire de choisir une valeur pour le support suffisamment basse afin d'extraire une information importante. Malheureusement, lorsque le seuil du support est trop bas, la quantité de règles extraites devient très grande rendant l'analyse de ces règles difficile. La confiance est une estimation de la précision des règles dans le futur. Cela représente la confiance désirée dans les règles.

Une certaine expertise est nécessaire afin de trouver les bonnes valeurs du support et de la confiance qui permettront d'obtenir les meilleures règles qui impactent la métrique F1. Pour trouver les valeurs optimales de ces deux paramètres, deux expérimentations ont été menées. Dans la première expérience, la valeur optimale du support a été recherchée en utilisant la

base del.icio.us. Nous avons fait varier le support sur un intervalle de 0.1 à 1 et nous avons sélectionné la valeur donnant les meilleures performances. La figure 4 montre l'évolution de la métrique F1 par rapport à la valeur du support.

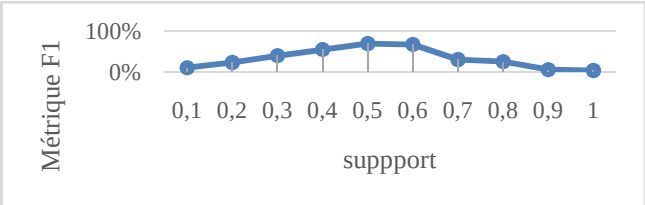


Figure 4 – Valeur optimale du support

Comme nous pouvons le voir sur cette figure, la meilleure valeur du support qui produit la plus grande valeur de la métrique F1 est 0.5. La seconde expérience concerne la recherche de la valeur optimale de la confiance en utilisant également la base del.icio.us pour un support minimal égal à 0.5. On a fait varier la confiance de la même façon que le support. La figure 5 montre l'évolution de la valeur de la métrique F1 par rapport à la confiance.

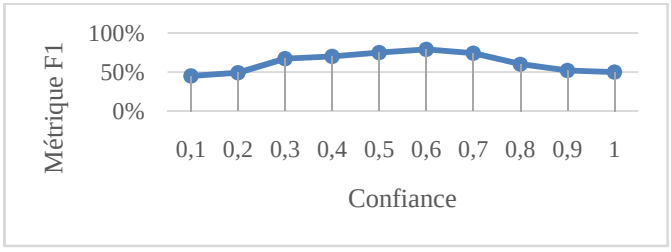


Figure 5 – Valeur optimal de la confiance

On en déduit que la valeur optimale de la confiance est 0.6. Il en résulte que des valeurs appropriées pour le support et la confiance sont respectivement 0.5 et 0.6.

4.4 Diversité dans la recommandation

Afin d'évaluer l'efficacité de notre approche pour donner des recommandations diversifiées, la métrique de diversité Intra-List proposée par (Ziegler et al., 09) est utilisée**.**

Dans cette section, nous évaluons la diversité de notre approche de recommandation et la comparons à la précision de celle-ci. Nous voulons ainsi voir si une augmentation de la diversité a bien lieu grâce à notre approche et si celle-ci a un impact sur la précision des recommandations. Nous avons testé notre approche sur trois niveaux d'exploration de LOD. Le nombre d'utilisateurs est égal à 20.

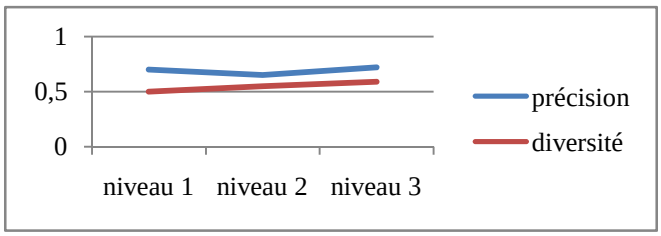


FIGURE 6 **–** Diversité Vs. Performance de la Recommandation

Les résultats présentés dans la figure 6 montrent que la recommandation basée sur les LOD augmente le taux de diversité avec le nombre d'items à recommander, provoquant juste une légère perte de précision.

La diversité dans notre approche est encore en cours d'évaluation et les premiers résultats montrent l'utilité d'explorer les LOD pour augmenter la diversité lors de la recommandation de ressources.

4.5 Passage à l'échelle

Les systèmes de recommandation étant destinés à aider les utilisateurs à naviguer dans de larges collections d'items, l'un de nos objectifs est de passer à l'échelle des Datasets réels. Il est donc important de mesurer la vitesse avec laquelle notre approche fournit des recommandations. Dans cette sous-section, on discutera de l'impact qu'à l'augmentation du nombre d'utilisateurs sur le temps d'exécution de notre approche. Afin de montrer la scalabilité de notre approche, nous avons mesuré le temps d'exécution requis afin de faire des recommandations dans la base del.icio.us avec un nombre d'utilisateurs allant de 1000 à 11500. La figure 7 montre que le temps d'exécution (en secondes) croît linéairement avec l'augmentation de la taille de la base de données. Cela signifie que notre approche répond bien à ce problème puisque l'augmentation du nombre d'utilisateurs dans la base de données provoquera approximativement une augmentation linéaire du temps d'exécution.

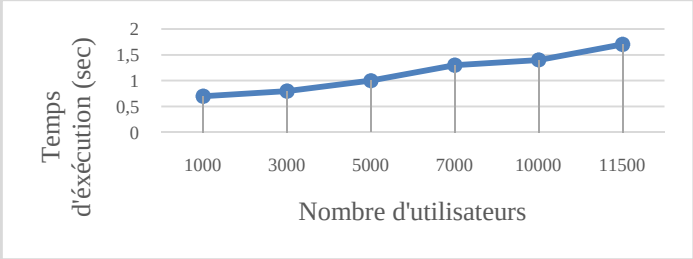


FIGURE 7 – Performance de notre approche lorsque la taille la base de données croît

5 Conclusion

Dans cette contribution, nous avons exploité la force de l'aspect social des folksonomies afin de permettre à chaque utilisateur de la communauté de bénéficier des ressources taguées par ses voisins dans le réseau social basé sur la recommandation de ressources. Nous avons vu l'importance d'analyser le profil de l'utilisateur afin de réaliser une recommandation dynamique et par conséquent l'importance de venir à bout des problèmes sémantiques inhérents aux folksonomies durant la recommandation. La méthode suivie est basée sur la similarité entre les utilisateurs dans certains cas et entre ressources LOD dans d'autres cas afin de venir à bout du problème du démarrage à froid lors de la recommandation. Les premiers résultats montrent l'intérêt d'explorer le graphe des LOD afin d'assurer la diversité lors de la recommandation de ressources personnalisées dans les systèmes d'étiquetage social.

Références

S. S. Anand, P. Kearney, et M. Shapcott. Generating semantically enriched user profiles for web personalization. ACM Trans. Internet Technol., 7(4), Oct. 2007.

V. A. A. Ayala, M. Przyjaciel-Zablocki, T. Hornung, A. Schatzle, et G. Lausen. Extending sparql for recommendations. In Proceedings of Semantic Web Information Management on Semantic Web Information Management, SWIM'14, pages 1:1-1:8, New York, NY, USA, 2014. ACM.

S. Beldjoudi, H. Seridi et C. Faron-Zucker. Ambiguity in Tagging and the Community Effect in Researching Relevant Resources in Folksonomies. In Proc. of ESWC workshop User Profile Data on the Social Semantic Web, 2011.

S. Beldjoudi, H. Seridi et C. Faron-Zucker. Improving Tag-based Resource Recommendation with Association Rules on Folksonomies. In Proc. Of ISWC workshop on Semantic Personalized Information Management: Retrieval and Recommendation, 2011.
S. Beldjoudi, H. Seridi et C. Faron-Zucker. Personalizing and Improving Tag-based Search in Folksonomies. In Proc. Of the 15th International Conference on Artificial Intelligence Methodology, Systems, Applications (AIMSA), Springer LNAI 7557, pp. 112-118, 2012.
P. De Meo, G. Quattrone, et D. Ursino. A query expansion and user profile enrichment approach to improve the performance of recommender systems operating on a folksonomy. User Modeling and User-Adapted Interaction, 2010.
M. Dojchinovski et T. Vitvar. Personalised access to linked data. In EKAW, pages 121-136, 2014.
C.L Huang, H.Y Chien, et M Conyette, Folksonomy-based Recommender Systems with User.s Recent Preferences, World Academy of Science,Engineering and Technology 78, 2011.
H. Khrouf et R. Troncy. Hybrid event recommendation using linked data and user diversity. In Proceedings of the 7th ACM Conference on Recommender Systems, RecSys '13, pages 185-192, New York, NY, USA, 2013.
N. Marie, O. Corby, F. Gandon, et M. Ribiere. Composite interests' exploration thanks to on-they linked data spreading activation. In Proceedings of the 24th ACM Conference on Hypertext and Social Media, HT '13, pages 31-40, New York, NY, USA, 2013.
S. E. Middleton, N. R. Shadbolt, et D. C. De Roure. Ontological user profiling in recommender systems. ACM Trans. Inf. Syst., 22:54-88, January, 2004.
P. Mika. Ontologies are us: A unified model of social networks and semantics. In Proc. of 4th Int. Semantic Web Conference (ISWC 2005), Galway, Ireland, volume 3729 of LNCS. Springer, 2005.
B. Mobasher, X. Jin, et Y. Zhou. Semantically enhanced collaborative filtering on the web. In B. Berendt, A. Hotho, D. Mladenic, M. Someren, M. Spiliopoulou, et G. Stumme, editors, Web Mining: From Web to Semantic Web, volume 3209 of Lecture Notes in Computer Science, pages 57-76. Springer Berlin Heidelberg, 2004.
M. Rowe. Semanticsvd++: incorporating semantic taste evolution for predicting ratings. In 2014 IEEE/WIC/ACM International Conferences on Web Intelligence, WI 2014, 2014.
M. Rowe. Transferring semantic categories with vertex kernels: recommendations with semanticsvd++. In The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, 2014.
V. Zanardi, L. Capra. A Scalable Tag-based Recommender System for New Users of the Social Web. In: Proc. of the 2nd International Conference on Database and Expert Systems Applications, 2011.
C. Ziegler, G. Lausen et G. Georges-Köhler-Allee. Making Product Recommendations More Diverse. IEEE Data Eng. Bull., 32(4), pp. 23-32, 2009.

Recommandation argumentée de ressources pédagogiques au sein d'un écosystème apprenant

Chahrazed Mediani, Marie-Hélène Abel

[1] Département d'Informatique, Université Ferhat Abbas de Sétif -1-,
Laboratoire des réseaux et des systèmes distribués, Sétif, Algérie,
Chahrazed_mediani@yahoo.fr
[2] Sorbonne universités, Université de technologie de Compiègne
UMR CNRS 7253, laboratoire HEUDIASYC, CS 60319,
60203 Compiègne Cedex, France,
marie-helene.abel@utc.fr

Résumé : Un écosystème apprenant est un ensemble cohérent composé de biocènes de formation favorisant un « apprentissage ensemble » basé sur l'échange et le partage de connaissances et/ou de compétences pour mieux réussir un projet commun. Les biocènes de formation sont variés et profitent de l'avancée des technologies de l'information et de la communication. Ainsi, la plateforme MEMORAe est un environnement de collaboration organisé autour d'un ensemble d'écosystèmes apprenants partageant des ressources pédagogiques issues de biocènes de formation. Bien que partagées, ces ressources peuvent présenter un degré de pertinence plus ou moins fort selon les membres de l'écosystème et l'objectif visé. Dans cet article, nous proposons une approche de recommandation argumentée de ressources pédagogiques au sein d'un écosystème apprenant. Cette approche est basée sur un système de vote permettant à chaque membre de l'écosystème de juger la pertinence d'une ou de plusieurs ressources pédagogiques se trouvant à sa disposition dans son écosystème apprenant.

Mots-clés : Ecosystème apprenant, système de vote, système de recommandation, ressource pédagogique.

1 Introduction

Dans le contexte des Technologies de l'Information et de la Communication (TIC), et des technologies du web 2.0, le terme d'écosystème est utilisé pour désigner l'ensemble des entités qui interagissent dans un environnement technologique (Ficheman et al., 2008). Dans le monde de l'entreprise, un écosystème est un ensemble composé d'entités (organisation, entreprises) et de leurs parties prenantes (client, fournisseurs, employeurs, etc.) ayant en commun un projet de développement où chaque partie s'engage à honorer des engagements prédéfinis vis-à-vis des autres. Côté formation, les écosystèmes apprenants ont pour but de faciliter un apprentissage basé sur les échanges et le partage de ressources pédagogiques entre apprenants. Ils sont généralement organisés autour d'un ensemble d'espaces de collaboration dédiés aux apprenants travaillant en groupe sur un même problème (Pettersson et al., 2010), formant ainsi un lieu de travail et d'échange pour le groupe, et permettant à chaque membre du groupe d'accéder aux ressources (documents et autres) destinées au groupe. L'apprentissage mis en œuvre dans un écosystème apprenant peut être qualifié d'apprentissage ensemble. Il permet à un apprenant de construire ses connaissances à partir d'interactions avec son entourage et son environnement (documents, échanges, etc.). Lorsque l'écosystème apprenant exploite les TIC pour sa mise en œuvre, les espaces de partage constituent une mémoire de travail (documents, idées, connaissances, solutions, etc.) relatif au problème traité. Les interactions sont à l'origine de la création et/ou l'identification de ressources pédagogiques au sein de ces espaces de partage (partage d'un document, réponse à

une question, annotation d'un document, etc.). Ces ressources peuvent être jugées plus ou moins pertinentes selon l'apprenant et l'objectif visé. Les systèmes de recommandations ont pour objectif de proposer des ressources jugées pertinentes à un utilisateur en se basant sur des informations sur les utilisateurs et les ressources du système. Nous nous intéressons dans notre recherche aux systèmes de recommandation de ressources pédagogiques dans les Environnements Informatiques Pour l'Apprentissage Humain (EIAH), et plus particulièrement dans un écosystème apprenant qui se base sur les interactions sociales entre apprenants, afin de les guider dans leur apprentissage.

Dans la suite, nous énonçons notre problématique avant de présenter des travaux existants liés aux systèmes de recommandations auprès des apprenants. Nous détaillons alors notre approche de recommandation argumentée basée sur un système de vote sur la pertinence de ressources au sein d'une communauté d'apprenants. Nous illustrons nos travaux au sein de la plateforme MEMORAe en nous bansant le modèle sous-jacent memorae-core2 que nous étendons. Nous concluons avant d'avancer des perspectives à ce travail.

2 Motivations

Un écosystème apprenant est un ensemble cohérent composé de biocènes de formation favorisant un « apprentissage ensemble » basé sur l'échange et le partage de connaissances et/ou de compétences pour mieux réussir un projet commun. Une biocène de formation est un moyen d'interagir avec le biotope ou environnement dans lequel les apprenants évoluent. Les biocènes de formation sont variées et profitent de l'avancée des technologies de l'information et de la communication. Au sein d'un écosystème apprenants, les apprenants peuvent être confrontés au partage de nombreuses ressources pédagogiques issues des différentes biocènes. Devant un nombre important de ressources pédagogiques, les apprenants peuvent se trouver perdus : quelles ressources consulter en priorité ? Les apprenants peuvent donc éprouver le besoin d'être aidés par des services leur permettant de choisir quelles ressources accéder pour atteindre leur objectif. Dans le cadre de nos travaux, nous nous intéressons à l'usage de systèmes de recommandation au sein d'écosystèmes apprenants afin d'aider les apprenants à accéder aux ressources pédagogiques servant leur objectif. Une ressource pédagogique peut être très variée selon la biocène utilisée : un site web, un livre, un support de cours, un jeu sérieux, une personne, une réponse à une question, etc. Elle peut être la trace d'une interaction (qui a fait quoi, quand, pourquoi et où) ou la production de cette interaction (la question posée, la réponse donnée, le document partagé, etc.). Nous avons choisi de développer notre système de recommandation en exploitant un modèle sémantique définissant un écosystème apprenant. Ce modèle doit prendre en considération un modèle des ressources pédagogiques, un modèle de l'apprenant, un modèle des traces d'interaction ainsi qu'un modèle de collaboration.

Afin de développer nos travaux, nous avons repris ceux effectués dans le cadre de l'approche MEMORAe. Cette dernière a permis de modéliser et concevoir une plateforme web permettant de gérer l'ensemble des ressources hétérogènes de connaissances circulant dans une organisation.

La plateforme, du même nom que l'approche, a été pensée et développée afin de faciliter l'apprentissage organisationnel et la capitalisation des connaissances à partir d'une modélisation sémantique : le modèle de collaboration memorae-core2. Elle exploite la puissance des nouvelles technologies support à la collaboration (technologies web 2.0.) et s'appuie sur les standards du web sémantique. Au sein d'un tel modèle, il est précisé QUI collabore sur QUOI, COMMENT et POURQUOI.

Le cœur de l'innovation concerne l'organisation autour d'une carte de connaissances de l'ensemble des ressources privées ou partagées, issues d'un processus formel ou informel au sein d'un groupe d'individus (équipe, service, projet, organisation, etc.). L'usage d'une carte sémantique permet de définir un référentiel commun dans lequel il est possible de

naviguer pour accéder aux ressources capitalisées dans différents espaces. Ces espaces sont visibles en parallèle et facilitent le transfert de connaissances entre individus.

Le modèle memorae-core2 fait la distinction entre les acteurs du monde et les utilisateurs de la plateforme. Il permet ainsi de modéliser les différentes interactions pouvant être réalisées au sein de la plateforme : homme-système ou bien homme-homme via le système. Il permet également de modéliser le rôle des personnes qu'elles soient ou non des utilisateurs de la plateforme.

En ce sens, la plateforme MEMORAe peut être considérée comme un écosystème apprenant numérique voire plusieurs écosystèmes liés par le partage d'un référenciel commun.

Afin de mettre en place notre système de recommandation, nous avons commencé par raffiner le modèle de trace de collaboration proposé par (Wang et al, 2014) et validé par la plateforme MEMORAe (Abel, 2009) avec un certain nombre d'indicateurs d'apprentissage décrivant l'état des activités de l'apprenant et la progression de ses connaissances lorsqu'il interagit au sein d'une communauté d'apprenants. En fonction de ces indicateurs, des recommandations sont données à l'apprenant (Mediani et al, 2015). Ces recommandations consistent en des actions à effectuer avec des membres de sa communauté jugés experts dans le domaine de sa tâche.

Dans le cadre de cet article, nous nous intéressons à un autre type de recommandation à savoir : la recommandation des ressources pédagogiques jugées pertinentes par les apprenants de l'écosystème en utilisant un système de vote.

3 Travaux connexes

Les systèmes de recommandation ont pour objectif de générer des suggestions sur de nouveaux sujets ou de prédire l'utilité d'un sujet pour un utilisateur donné. Parmi les travaux qui ont été menés dans le contexte des EIAH afin de générer des recommandations pour soutenir les apprenants dans leur apprentissage, nous pouvons citer les travaux de (Peis et al., 2008) qui utilisent une approche basée sur le contenu : on recommande à l'utilisateur des sujets similaires à ceux qu'il a déjà appréciés. D'autres approches de recommandations collaboratives ont été utilisées. Elles sont basées sur les appréciations d'un ensemble d'utilisateurs sur les items et on distingue les approches basées sur les items (Sarwar, 2001) et celles basées sur les utilisateurs (Resnick et al, 1994). Des approches hybrides ont été aussi utilisées. Elles combinent de différentes manières les approches précédentes : (Berkani et al. 2012) combinent les approches de filtrage à base de contenu et collaboratif pour la recommandation personnalisée de ressources pédagogiques dans une Communauté de Pratique de E-learning. Dans le cadre de notre approche Memorae, (Li et al., 2012) définissent un modèle de trace original qui distingue les actions privées, individuelles, collectives et collaboratives. (Wang et al. 2014) définissent une méthode d'exploitation de ce modèle basé sur la méthode des TF-IDF pour calculer l'indice de compétence de chaque apprenant concernant un élément de connaissance donné et pour proposer un système de recommandation exploitant cet indice de compétence. (Mediani et al, 2015) ont raffiné le modèle de trace de Memorae par un certain nombre d'indicateurs d'apprentissage décrivant l'état des activités de l'apprenant et la progression de ses connaissances dans sa communauté d'apprenants pour suggérer des recommandations à l'apprenant. Ces recommandations consistent à suggérer à l'apprenant d'effectuer des activités avec des membres de sa communauté jugés experts dans le domaine de sa tâche.

4 La plateforme MEMORAe et la recommandation

La plateforme MEMORAe est un environnement de collaboration qui permet à plusieurs utilisateurs d'interagir simultanément avec l'environnement pour accéder aux concepts d'un cours et aux ressources indexées par ces concepts définis au sein d'une carte de connaissance

(Abel et Leblanc, 2009). L'environnement permet à chaque utilisateur de choisir d'accéder à différents espaces de partage de ressources : l'espace privé ou les espaces de groupes auxquels il appartient.

- L'espace privé : espace où l'apprenant peut référencer ses propres ressources. Le contenu de cet espace n'est accessible que par cet utilisateur.
- Espace de groupe : espace uniquement accessible par les membres du groupe associé. C'est au sein de cet espace que les membres partagent et échangent des ressources.

4.1 MEMORAe comme écosystème apprenant

L'écosystème apprenant MEMORAe a pour objectif d'aider l'apprenant utilisateur à appréhender les concepts d'une formation et à faciliter les échanges et le transfert de connaissances au sein de l'écosystème autour de ces concepts. Dans ce cadre, le contenu pédagogique d'une formation est composé d'une ontologie d'application et des ressources indexées par les concepts de cette dernière. Les concepts de l'ontologie d'application définissent formellement les notions à appréhender et permettent d'indexer les ressources pédagogiques les traitant contribuant ainsi à leur appréhension. Afin de mieux guider l'apprenant, nous avons fait le choix d'associer un attribut poids aux liens de spécialisation de la taxonomie de l'ontologie d'application. Un attribut poids permet de préciser le degré de contribution d'un concept dans l'appréhension de son concept père. La valeur de l'attribut poids est comprise entre 0 et 1, la somme des attributs poids associés aux liens de spécialisation d'un concept est quant à elle égale à 1. La Figure 1 illustre les poids définis dans le cadre de l'ontologie d'application du cours « Information Technology » enseigné à l'université de Sétif.

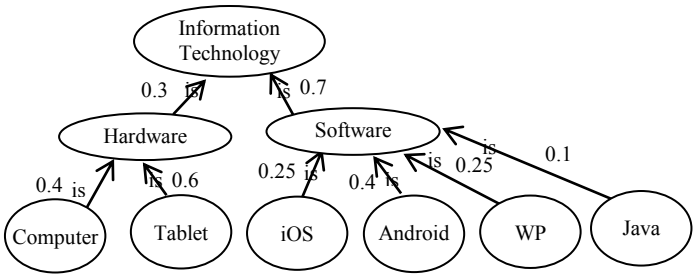


Figure 1 –Extrait de l'ontologie d'application « Information Technology » (Mediani et al., 2015).

Les activités de l'apprenant réalisées au sein de l'écosystème numérique MEMORAe sont sauvegardées dans une base de traces et classifiées en trois types de traces : (1) les traces privées enregistrent les activités que l'apprenant effectue dans son espace privé ; (2) les traces de collaboration enregistrent les activités effectuées par les apprenants dans leur espace de collaboration ; (3) les traces individuelles enregistrent les activités effectuées par l'apprenant dans ses espace privés et de collaboration. Pour chaque type de traces, nous avons trois types d'activités qui peuvent être menées par l'apprenant : les consultations de ressources pédagogiques (les documents), les créations de ressources (les conversations, les meetings, les questions, les réponses, les notes et les wikis) et les ajouts de ressources (documents et annotations).

Exemple : La Table 1 récapitule les activités réalisées par les apprenants de l'écosystème « groupe 1 » et concernant les différents concepts de la Figure 1. Pour un concept donné, chaque cellule du tableau représente le nombre d'activités réalisées par l'apprenant (C : Consultation, R : Création, A : Addition). Le nombre avant la parenthèse est la somme totale des activités réalisées par l'apprenant concernant ce concept.

Table 1. Tableau récapitulatif des actions des apprenants du groupe 1.

	Elsa	Jean-Paul	Ning	Marie-Hélène	Total
Java	0(0C,0R,0A)	0(0C,0R,0A)	0(0C,0R,0A)	0(0C,0R,0A)	**0(0C,0R,0A)**
WP	0(0C,0R,0A)	2(1C,1R,0A)	1(0C,0R,1A)	2(0C,1R,1A)	**5(1C,2R,2A)**
Android	4(1C,3R,0A)	8(4C,3R,1A)	3(1C,1R,1A)	5(4C,1R,0A)	**20(10C,8R,2A)**
Ios	5(2C,1R,2A)	0(0C,0R,0A)	5(2C,2R,1A)	1(0C,0R,1A)	**11(4C,3R,4A)**
Tablet	0(0C,0R,0A)	1(0C,1R,0A)	2(1C,0R,1A)	3(0C,2R,1A)	**6(1C,3R,2A)**
Computer	3(2C,0R,1A)	0(0C,0R,0A)	2(0C,0R,2A)	0(0C,0R,0A)	**5(2C,0R,3A)**
Software	4(2C,1R,1A)	0(0C,0R,0A)	4(1C,2R,1A)	3(1C,1R,1A)	**11(4C,4R,3A)**
Hardware	2(1C,1R,0A)	1(0C,1R,0A)	2(1C,0R,1A)	3(0C,2R,1A)	**8(2C,4R,2A)**
Info_Tech	1(1C,0R,0A)	1(1C,0R,0A)	1(1C,0R,0A)	1(1C,0R,0A)	**4(4C,0R,0A)**
Total	**19(9C,6R,4A)**	**13(6C,6R,1A)**	**20(7C,5R,8A)**	**18(6C,7R,5A)**	**60(28C,24R,18A)**

C : Consultation, **R** : Création, **A** : Addition.

4.2 Recommandation d'activités

Dans des travaux précédents (Mediani et al., 2015), nous avons raffiné le modèle de trace de collaboration de (Li et al., 2012), et repris par (Wang et al., 2014), par un certain nombre de mesures permettant de construire des indicateurs sur l'état de connaissances de l'apprenant et sur la progression de ses connaissances au sein d'un groupe dans une session d'apprentissage. Parmi ces paramètres, nous avons retenu le degré de maîtrise d'une connaissance représentée par un concept. Pour atteindre ces objectifs, nous avons adopté la démarche suivante : (i) proposer un modèle sémantique pour mesurer des indicateurs de la contribution de chaque apprenant au sein de son groupe, (ii) estimer ces indicateurs de contribution en prenant en compte les connaissances de l'apprenant ainsi que ses activités, (iii) proposer un ensemble de recommandations pour aider l'apprenant dans son apprentissage et le préparer pour une évaluation plus pertinente. On recommande à l'apprenant d'effectuer certaines activités avec des membres de son écosystème jugés experts dans le domaine de sa tâche. Pour mesurer ces indicateurs d'apprentissage, nous avons considéré que la maîtrise d'un concept C est liée aux activités réalisées, par l'apprenant dans son écosystème, concernant le concept C et les concepts SC spécialisant C. Pour chaque concept C, nous distinguons deux poids : P1 pour les activités concernant directement C et P2 pour les activités concernant les concepts SC spécialisant C. La somme des deux poids doit être égale à un. **Exemple :** P1= 0.6, P2 = 0.4.

Nous avons associé également à chaque type d'activité un poids pris en compte dans le calcul du degré de maîtrise de ce concept par l'apprenant. La somme des poids des types d'activités doit également être égale à un.

Exemple : Pour un concept C, le poids de consultation PC(C) = 0.2, le poids de création PR(C) = 0.5 et le poids d'addition PA(C)=0.3. Ces poids peuvent varier d'un concept à l'autre.

4.3 La recommandation de ressources pédagogiques

Au sein de la plateforme MEMORAe il est possible de partager des ressources pédagogiques hétérogènes avec les membres d'un écosystème apprenant. Ces ressources peuvent présenter pour un même sujet un degré de pertinence variable. Ainsi un professeur faisant des groupes de niveau pour son cours pourra recommander un même tutoriel pour un même sujet avec un degré de pertinence différent selon le groupe de niveau : fortement recommandé pour un groupe faible et faiblement recommandé pour un groupe plus fort. Cette différence de recommandation devra être mise en œuvre au sein des espaces de partage associés aux deux écosystèmes. La ressource recommandée peut être de différentes formes comme par exemple la réponse à une question posée au sein de l'écosystème dans le cadre du

forum accessible dans l'espace de partage. Notre objectif est d'aider les apprenants dans leur choix de consultation et/ou les amener à se prononcer sur l'intérêt d'une ressource dans son écosystème. Il s'agit de construire de manière participative des éléments de réponse de façon à déterminer les ressources les plus pertinentes pour le concept qui les indexe et les plus adaptées au profil de l'utilisateur pour appréhender ce concept.

4.3.1 Le système de vote

Dans cette section, nous présentons la mise en œuvre d'un système de vote participatif dans le cadre d'un écosystème apprenant MEMORAe permettant de calculer un degré de pertinence d'une ressource pour un sujet/concept au sein de l'écosystème. Ce sont les membres d'un écosystème apprenant qui peuvent attribuer un degré de pertinence à une ressource partagée pour juger de sa pertinence pour l'appréhension du concept qui l'indexe. Ce degré peut prendre une valeur variant de 1 à 5 (1 : fortement recommandé, 2 : recommandé, 3 : indécis, 4 : faiblement recommandé, 5 : non recommandé). Le degré de pertinence final sera calculé en faisant la moyenne de tous les votes.

Dans le modèle memorae-core2 que nous avons étendu (cf. Figure 2), un vote est une ressource simple qui a pour cible un *indexkey*. Un *indexkey* est un concept faisant le lien entre une ressource pédagogique, un concept qui indexe cette ressource et un espace de partage où cette ressource est visible/partagée. De la sorte, il devient possible de voter pour le degré de pertinence d'une ressource pour une communauté (espace de partage) pour un concept particulier. La pertinence pourra changer en fonction du concept et cela même dans le même espace. Le livre L1 est très pertinent pour le concept C1 et moins pour le concept C2 et cela pour la même communauté. Le fait de voter se traduira par la création d'une ressource simple vote V1 par l'utilisateur U1 au sein de la plateforme ayant pour cible un indexkey K1. K1 fait le lien avec l'espace S1, le concept concerné et la ressource R1 évaluée.

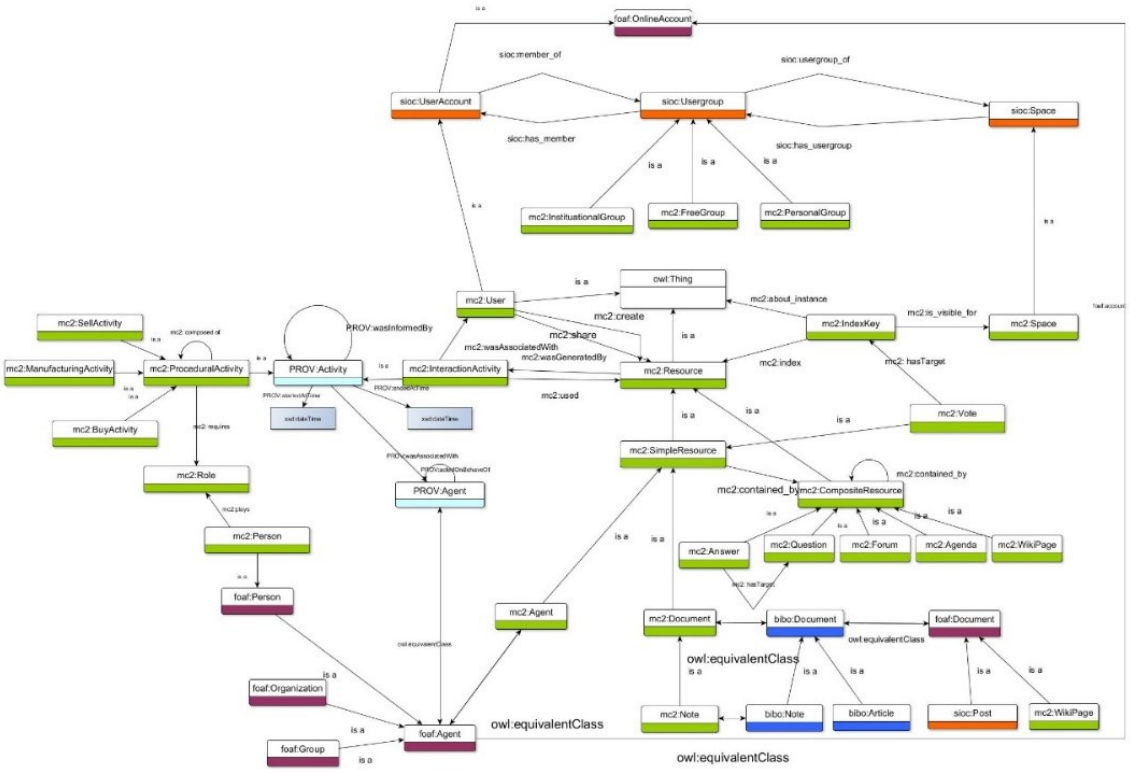


FIGURE 1 – Le modèle de memorae-core2

Concernant l'interface, nous avons ajouté un bouton de vote dans la fenêtre qui affiche les informations sur la ressource pédagogique (cf. Figure 2), mentionnant si l'utilisateur a voté ou non comme celui des annotations. La Figure 2 concerne le vote pour une ressource de type documentaire. Un clic sur le bouton Vote permet à l'utilisateur de :

- Voter sur la pertinence de la ressource en attribuant une valeur entre 1 et 5 si l'utilisateur n'a pas encore voté sur cette ressource.
- Modifier la valeur de son vote s'il a déjà voté et s'il veut changer d'avis.
- Visualiser la moyenne de vote, pour cette ressource, de tous les utilisateurs de cet espace de collaboration.

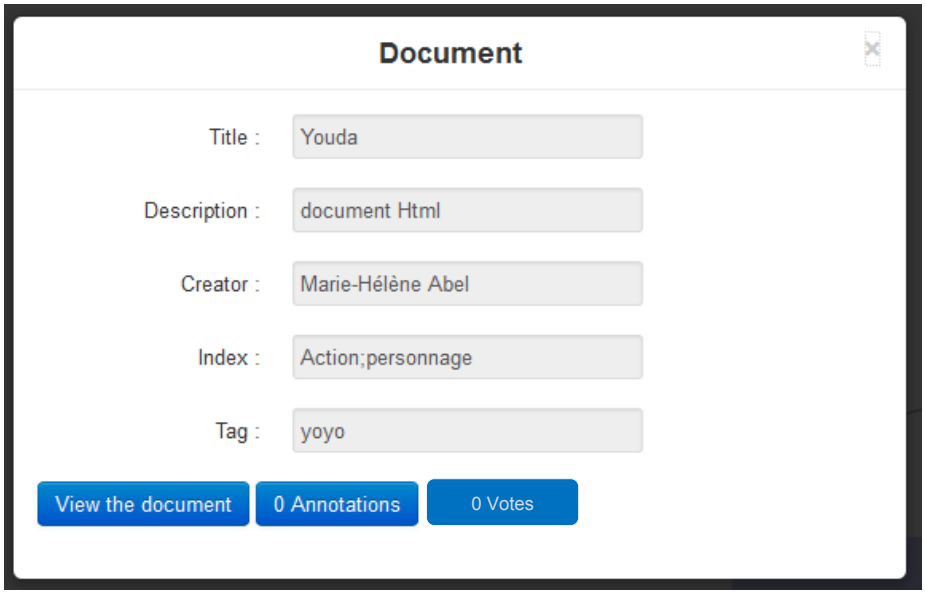


FIGURE 2 – Une ressource pédagogique de type « Document »

4.4 Le système de recommandation de ressources pédagogiques

Le système de recommandation que nous proposons est basé sur deux modèles : le modèle de l'apprenant/utilisateur et le modèle de contenu pédagogique et plus particulièrement la ressource pédagogique. Dans cette section, nous présentons brièvement les deux modèles ensuite nous présentons l'approche de recommandation de ressources pédagogiques.

4.4.2 Le modèle de l'apprenant

Le contenu pédagogique est décomposé en un ensemble d'éléments et le modèle de l'apprenant est représenté par un ensemble de valeurs mesurables associées à ces éléments. Ces valeurs varient entre 0 (non maîtrisé) et 1 (maîtrisé). Dans notre modèle, nous considérons que la maîtrise d'un concept C est liée aux activités réalisées concernant C et aussi concernant les sous concepts SC spécialisant C. Parmi ces indicateurs que nous avons utilisés dans (Mediani et al, 2015) pour mesurer le degré de maitrise d'un concept : la contribution par les activités AC(l, c, s), la contribution par les sous-concept KC(l, c, s) et la contribution globale (degré de maîtrise) KL(l, c, s) où l : représente l'apprenant et c : est un concept donné et s : est l'espace de partage.

Dans (Mediani et al., 2015), nous avons montré comment calculer les contributions des apprenants pour chaque concept de l'ontologie d'application de la Figure 1.

4.4.2.1 La contribution par les activités.

Pour un apprenant l dans un espace s, l'indice de la contribution par activité AC(l, c, s) pour un concept c est calculé comme suit :

$$\mathrm{AC(l,c,s)} = \sum_{k=1}^{n} P_c(k) * contribution\_value_{(l,c,s)}(\mathrm{k}) \quad (1)$$

Avec n : le nombre des types d'activités, dans notre cas n=3, et $P_c(k)$: le poids du type de l'activité k (consultation, création ou addition) pour le concept c.

contribution_value(l, c, s)(k) est une fréquence relative estimée par le rapport entre le nombre d'activités de type k, concernant le concept c, réalisées par l'apprenant l au sein de son groupe s et le nombre de toutes les activités de type k, concernant le même concept c, réalisées par l'ensemble des membres du groupe s. contribution_value est soit consultation_value, creation_value ou addition_value.

Exemple : Pour le concept « Android », supposons que : PC(Android)=0.2, PR(Android)=0.5 et PA(Android)=0.3. En utilisant la table 1, Marie-Hélène a effectué 4 activités de consultation, 1 création et 0 ajout dans le groupe 1. Donc :

Consultaion_value(Marie-Hélène, Android, S1)= 4/10 = 0,400

Creation_value(Marie-Hélène, Android, S1)= 1/8 = 0,125

Addition_value(Marie-Hélène, Android, S1)= 0/10 = 0,000

Nous calculons donc la contribution par les activités de Marie-Hélène du groupe 1 pour le concept « Android » comme suit :

AC(Marie-Hélène, Android, S1) = 0.400*0.20+0.125*0.5+0.000*0.3 = 0.142

4.4.2.2 La contribution par les sous-concepts.

Pour un apprenant l de l'espace s, l'indice de la contribution par les sous-concepts k pour un concept c KC(l, c, s) est égal à :

$$\mathrm{KC(l,c,s)} = \sum_{k=1}^{n} P(k) * \mathrm{KL(l,k,s)} \quad (2)$$

n est le nombre des sous-concepts k du concept père c. P(k) : le poids associé à chaque sous concept k (ces poids sont définis dans l'ontologie d'application). KL(l, k, s) est le niveau de connaissance de l'apprenant l sur le concept k dans l'espace s.

Exemple : Le concept « Android » n'a pas de sous-concept donc :

KC(Marie-Hélène, Android, S1) = 0

Si KL(Marie-Hélène, Ios, S1) = 0,075, KL(Marie-Hélène, Android, S1) = 0,142, KL(Marie-Hélène, WP, S1) = 0,400 et KL(Marie-Hélène, Java, S1) = 0,000, alors en appliquant l'équation (2) :

KC(Marie-Hélène, Software, S1) = 0.25*0.075+0.4*0.142+0.25*0.400+0.1*0.000= 0.159

4.4.2.3 La contribution globale (degré de maîtrise)

Le degré de maîtrise ou le niveau de connaissance de l'apprenant l sur le concept c KL(l, c, s) est égal à :

$$KL(l, c, s) = P1 * AC(l, c, s) + P2 * KC(l, c, s) \quad (3)$$

P1 et P2 sont les poids associés aux deux contributions (*activities_contribution* et *knowledge_contribution*).

Exemple : Nous allons inférer le degré de maîtrise de Marie-Hélène pour le concept « Software ». Supposons, pour ce concept, que le poids P1 associé aux activités est égal à 0.6 et le poids P2 associé aux sous-concepts est égal à 0.4.

KL(Marie-Hélène, Software, S1)= P1 * AC(Marie-Hélène, Software, S1)+ P2 * KC(Marie-Hélène, Software, S1)

En appliquant l'équation (3) :

KL(Marie-Hélène, Software, S1) = 0.6*0.274+0.4*0.159=0.228

4.4.3 Le modèle de la ressource

Dans notre modèle de contenu pédagogique, une ressource est décrite par un ensemble de métadonnées : le titre, la description, le type, l'auteur, la langue, le format, la date de création, la date de modification, le domaine de connaissance, les mots clés et la difficulté. La ressource est indexée par les concepts de la formation et elle a un niveau de difficulté (faible, moyenne et haute). Ce niveau de difficulté est attribué par l'utilisateur qui a ajouté la ressource dans l'espace de partage et il est utilisé dans le filtrage de ressources selon le degré d'expertise de l'utilisateur. Dans notre système de recommandation, nous recommandons à l'utilisateur la ressource ayant le niveau de difficulté adapté à son degré de maîtrise.

Ici par exemple, nous montrons comment deux ressources « course1» et «course2 » sont indexées par le même concept «Software», ayant le même niveau de difficulté (faible) et sont visibles dans le même espace de partage S1 du groupe 1. Ainsi, ces deux ressources ont deux *IndexKeys* Ik1 et IK2 respectueusement comme il est indiqué dans le tableau suivant :

Table 2. Tableau montrant deux IndexKeys différents.

Ik1:	Ik2:
about_instance: Software; visible_for: S1; index : course1.	about_instance: Software; visible_for: S1; index : course2.

Le 10 janvier 2016, **Elsa** a jugé « course1» très pertinent pour appréhender le concept «Software» et a attribué un vote de valeur 1 au sein de l'espace de partage S1. En même temps, **Ning**, dans le même espace S1, n'était pas d'accord et il a voté 3 pour la même ressource. En outre, **Elsa** a attribué un vote de valeur 4 à la ressource « course2 » indexée par le même concept "Software" visible dans S1 car, selon elle, elle est beaucoup moins pertinente. Après ces trois activités, nous obtenons trois ressources de votes, comme il est indiqué dans le tableau Table 2 :

Table 2. Tableau montrant trois votes différents

Vote_1:	Vote_2:	Vote_3:
Creator: Elsa	Creator: Ning	Creator: Elsa
Value_of_vote: 1	Value_of_vote: 3	Value_of_vote: 4
IndexKey; Ik1	IndexKey; Ik1	IndexKey; Ik2
Date: 10-01-2016	Date: 10-01-2016	Date: 10-01-2016

Exemple : Dans l'exemple précédent, Marie-Hélène a un niveau de connaissance faible pour le concept « Software » : KL(Marie-Hélène, Software, S1) = 0,228. Les deux ressources « course1» et « course2 » sont indexées par le même concept « Software », ayant le même niveau de difficulté (faible) et sont visibles dans le même espace de partage S1 du groupe 1. La ressource « course1» est recommandée à Marie-Hélène par le système de recommandation car elle a un niveau de difficulté faible et elle est mieux votée que la ressource « course2».

4.4.4 Le module de recommandation

Notre module de recommandation de ressources pédagogiques se base sur les liens sémantiques existant entre les concepts de la formation, les ressources indexées en tenant compte du degré de pertinence calculé, le niveau de difficulté qu'elles présentent et le modèle de l'apprenant. Ainsi pour un apprenant *ap* désirant appréhender une connaissance *c* ayant accès à l'espace de partage *s*, nous lui recommanderons des ressources du système qui sont :

- Indexées par le concept *c* et/ou les concepts *sc* spécialisant *c*.
- Accessible dans l'espace de partage de la communauté à laquelle appartient *ap*.
- Ayant une valeur de vote supérieure à un seuil (elle a été jugée pertinente par les membres de la communauté).
- Ayant une difficulté adaptée au degré de maitrise de l'apprenant *ap* pour le concept *c* qui indexe cette ressource.

Formellement une recommandation R consiste en une proposition d'une ou plusieurs ressources pédagogiques.

$$R = <l, c, s, (r_1, r_2, \ldots, r_n)>$$

- l : l'apprenant tracé.
- s : l'espace de travail.
- c : le concept concerné par la recommandation.
- $(r_1, r_2, \ldots, r_n)$: l'ensemble des ressources de l'espace s qui sont jugées pertinentes pour le concept c et sont adaptées au profil de l'apprenant.

L'algorithme suivi est :

```
Input : V: Averages votes of the pedagogical resources, l : Learner, s : Space, c :
Concept,votemin: minimum value of vote, RB : list of recommended resources.
Output : RB : list of recommended resources.
   KL := calculate_Knowledge_Level(l,c,s)
   RES:= search_resources(c,s)
   For all resource ri of RES do
    If (V(ri) > Votemin) then
       Adapted:=check_difficulty(ri,KL)
       If Adapted then
         Add(<l,c,s,ri>, RB)
       EndIf
    EndIf
   EndFor
   CO:=search_sub_concepts(c,s)
   For all sub-concept cj of CO do
    KL:=calculate_Knowledge_Leval(l,cj,s)
    If KL < ε then
      REs:= search_resources(cj,s)
      For all resource ri of RES do
        If (V(ri) > Votemin) then
```

```
        Adapted:=check_difficulty(ri,KL)
        If Adapted then
          Add(<l,cj,s,ri>, RB)
        EndIf
      EndIf
    EndFor
  EndIf
EndFor
End.
```

calculate_knowledge_level(l,c,s) : est une fonction qui calcule le niveau de connaissances (degré de maîtrise) de l'apprenant l sur le concept c visible dans l'espace de partage s. Sa valeur varie entre 0 (non maîtrisé) et 1 (maîtrisé).

search_ressouces(c,s) : est une fonction qui a pour résultat la liste des ressources pédagogiques indexées par le concept c visible dans l'espace de partage s.

search_sub_concept(c,s) : est une fonction qui a pour résultat la liste des sous concepts du concept c visible dans l'espace de partage s.

chech_difficulty(r,KL) : est une fonction qui vérifie si le niveau de difficulté KL de la ressource r est adapté au niveau de connaissances de l'apprenant. **Exemple:** une ressource de faible difficulté est adaptée à un faible niveau de connaissances (0 <= KL <= 1/3).

Pour une certaine ressource indexée par le concept c ou les concepts sc spécialisant c, lorsque la moyenne de vote de tous les utilisateurs de l'espace de partage est supérieure à un certain seuil et le niveau de difficulté de cette ressource est adapté au degré de maîtrise de l'utilisateur, on recommande cette ressource à l'apprenant.

Notre système de recommandation permet d'exploiter les indicateurs d'apprentissage décrivant l'état des activités de l'apprenant et la progression de ses connaissances dans sa communauté d'apprenants pour suggérer des recommandations à l'apprenant. Ces recommandations consistent à suggérer à l'apprenant des ressources jugées pertinentes par les membres de sa communauté et qui s'adaptent au mieux à son niveau de connaissance. Si par exemple l'apprenant a un niveau de connaissance moyen, on lui recommande des ressources pédagogiques de difficulté moyenne.

5 Conclusion

Dans le cadre de nos travaux nous nous intéressons aux écosystèmes apprenants. Un écosystème apprenant est un ensemble cohérent composé de biocènes de formation favorisant un « apprentissage ensemble » basé sur l'échange et le partage de connaissances et/ou de compétences pour mieux réussir un projet commun. Les biocènes de formation sont variés et profitent de l'avancée des technologies de l'information et de la communication. Ainsi, la plateforme MEMORAe est un environnement de collaboration organisé autour d'un ensemble d'écosystèmes apprenants partageant des ressources pédagogiques issues de biocènes de formation. Bien que partagées, ces ressources peuvent présenter un degré de pertinence plus ou moins fort selon les membres de l'écosystème et l'objectif visé. Dans cet article nous avons montré pourquoi et comment nous avons mis en place une aide au choix de consultation de ressources pédagogiques au sein d'un écosystème apprenant. Nous avons décidé d'exploiter une modélisation sémantique sur laquelle s'appuie un système de recommandation. La modélisation sémantique est sous-jacente à l'écosystème apprenant numérique que nous visons. Elle distingue principalement les ressources pédagogiques, les apprenants, la collaboration et permet de distinguer qui collabore sur quoi, pourquoi et comment. De façon à permettre d'argumenter la recommandation d'une ressource pédagogique au sein d'un écosystème, nous avons mis en place un système de vote qui permet de préciser l'intérêt d'une ressource selon un sujet et une communauté d'apprenants. Le

module de recommandation établi exploite les votes, la description de la ressource ainsi que le modèle de l'apprenant.
Ce travail a été réalisé au sein de la plateforme MEMORAe et a repris et étendu le modèle memorae-core2. Nous planifions de tester l'écosystème apprenant numérique obtenu auprès des étudiants de l'université de Sétif suivant le cours « Information Technology ». Une amélioration du module de recommandation en cours de réalisation concerne l'ajout d'une pondération sur un vote. Le vote d'un apprenant averti, selon le contexte, n'aura pas le même poids qu'un apprenant débutant. Afin d'offrir une argumentation plus fine, nous travaillons également à mieux exploiter le modèle de l'apprenant notamment en considérant les compétences et connaissances qu'il possède.

Références

Abel, M. H., Leblanc, A.: Knowledge Proc of sharing via the E-EMORAe2.0 platform. In: the International Conference on Intellectual Capital, Knowledge Management & Organizational Learning, (2009) 10-19.

Berkani, L., Nouali, O., Chikh, A. : Recommandation personnalisée des ressources dans une communauté de pratique de e-learning. Une approche à base de filtrage hybride. INFORSID, (2013) 131-138.

Ficheman, I.K., Lopes, R.D. : DIGITAL LEARNING ECOSYSTEMS: Authoring, Collaboration, Immersion and Mobility. IDC '08 Proceedings of the ACM 7th international conference on Interaction design and children, New York, NY, USA, (2008) 9-12.

Li, Q., Abel, M.H., Barthès, J.P.: Facilitating Experience Groups Sharing -Collaborative Trace. Proceeding of Reuse Exploitation. and In: International Conference on Knowledge Management and Information Sharing, (2012) 21-30.

Mediani, C., Abel, M. H., Djoudi, M.,: Towards a Recommendation System for the Learner from a Semantic Model of Knowledge in a Collaborative Environment. 5th IFIP TC 5 International Conference, CIIA 2015, Saida, Algeria, May 20-21, 2015, Proceedings. IFIP Advances in Information and Communication Technology 456, Springer 2015, ISBN 978-3-319-19577-3, (2015) 315-327.

Peis E., Morales-del-Castillo J. M., Delgado-López J. A. : Semantic Recommender Systems. Analysis of the state of the topic, Hipertext.net, n° 6, (2008).

Pettersson, O., Svensson, M., Gil, D., Andersson, J., Milrad, M. : On the Role of Software Process Modeling in Software Ecosystem Design. ECSA '10 Proceedings of the ACM Fourth European Conference on Software Architecture, New York, NY, USA, (2010) 103-110.

Resnick, P. Iacovou, N. Suchak, M. Bergstrom, P., Riedl, J. : Grouplens: an open architecture for collaborative filtering of netnews. Proceedings of the ACM conference on Computer supported cooperative work, New York, NY, USA, (1994)175–186.

Sarwar, B., Karypis, G., Konstan, J., Reidl, J. : Item-based collaborative filtering recommendation algorithms. Proceedings of the 10th international conference on World Wide Web, New York, NY, USA, (2001) 285–295.

Wang, N., Abel, M. H., Barthès, J.P., Negre, E.: Towards a Recommender System from Semantic Traces for Decision Aid. KMIS, Rome (2014) 274-279.

SemMEP : Nouvelle approche sémantique pour la détection des communautés dans un réseau social

Sami Ben Amor[1], Lotfi Ben Romdhane[1] et Mounira Harzallah[2]

[1] Groupe de recherche MARS, Université de Sousse,
ben_amor_sami@outlook.com, lotfi.ben.romdhane@gmail.com
[2] Data User Knowledge, LINA, Université de Nantes
mounira.harzallah@univ-nantes.fr

Résumé : Plusieurs travaux ont porté sur la détection des communautés dans les réseaux sociaux. La majorité d'entre eux considère seulement la structure d'un réseau en négligeant la richesse sémantique des informations associées à ses utilisateurs et aux liens entre eux. D'autres approches se sont focalisées sur ses aspects sémantiques. Récemment des nouvelles approches ont proposé une modélisation conjointe de ces deux aspects. Dans cet article, nous proposons une nouvelle approche favorisant l'aspect sémantique d'un réseau social tout en tenant compte de sa structure. Une nouvelle fonction de qualité et un nouvel algorithme SemMEP pour l'optimiser sont définis, utilisant les informations sémantiques d'un réseau dans plusieurs étapes du processus de détection, à l'aide des mesures sémantiques.

Mots-clés : Réseau social, analyse sémantique, ontologie, détection des communautés.

1 Introduction

Les réseaux sociaux, tels que Facebook, Twitter et Flickr, sont devenus un moyen de communication très important utilisé par le grand public ainsi que par les professionnels. En raison de l'explosion de ces réseaux et de la richesse de leur contenu, l'intérêt pour leur analyse a augmenté au cours des dernières années. Plusieurs travaux se sont intéressés à la détection des communautés dans ces réseaux. La majorité d'entre eux considère seulement la structure d'un réseau et néglige la richesse des informations liées à ses acteurs et aux liens entre eux. D'autres s'intéressent seulement à la sémantique de ces informations et ignorent l'aspect structurel d'un réseau. Pour faire face à ce problème, on a associé au lien entre deux acteurs un poids défini par une combinaison linaire de l'intensité structurelle entre eux et la similarité sémantique de leurs informations (Dang & Viennet, 2012) (Cruz et al. 2013) (Zhang et al. 2015). Cependant, ce poids n'a pas été évalué d'une façon homogène pour tous les couples d'acteurs. La majorité de ces approches applique des algorithmes classiques de détection des communautés dont le plus utilisé est celui de Louvain (Blondel et al. 2008).

Les informations dans ce type de réseaux ne sont pas faciles à prendre en compte pour la détection des communautés. En effet, on a besoin, en premier lieu, d'en extraire les plus pertinentes, d'identifier leur sémantique et de déterminer la similarité de celles associées à deux nœuds à l'aide des mesures de similarité. En deuxième lieu, on a besoin de savoir les intégrer dans une modélisation d'un réseau social et de bien les prendre en compte tout au long du processus de détection des communautés. Récemment, des nouvelles approches utilisent une ontologie pour définir la sémantique de ces informations. Erétéo et al. (2011) et Leprovost et al. (2012) considèrent seulement les concepts d'une ontologie qui généralisent ceux qui annotent un réseau social. Wan et al. (2014) attribuent seulement trois valeurs

possibles à la similarité sémantique des concepts qui annotent deux nœuds (1, 0.5, 0). Dans ces différentes approches, les aspects sémantiques d'un réseau social ne sont pas pris en compte que dans la première étape du processus.

Nous avons développé l'approche SemMEP proposant une nouvelle fonction de qualité « SemEP » mesurant la qualité d'un partitionnement du point de vue structurelle et sémantique. SemMEP, définie par analogie avec l'approche structurelle MWEP, utilise dans plusieurs étapes du processus de détection les résultats des mesures sémantiques appliquées à l'ontologie qui annote un réseau. Dans cet article, nous présentons notre approche SemMEP et son expérimentation sur le réseau de « Karaté » suivant plusieurs scénarios d'annotation.

2 SemMEP : Maximisation Sémantique de l'Equilibre et de la Pureté

Notre approche SemMEP est une extension de l'approche MWEP pour la détection des communautés dans un réseau social modélisé par un graphe pondéré (Zardi & Ben Romdhane, 2013b). MWEP a donné des bons résultats de la qualité de partitionnement et de la rapidité d'exécution. Dans cette approche, on a défini la fonction de qualité «WEP» et l'algorithme MWEP pour l'optimiser. Ce dernier est composé principalement de deux phases : la pureté et l'équilibre. La phase de pureté reflète l'attachement d'un nœud du réseau à une communauté en fonction de la force de ses relations avec ses voisins. Pour décider si un nœud v est pur par rapport à une communauté C_i et donc l'associer à cette communauté, on compare sa compatibilité à C_i, noté $comp(v, C_i)$ qui est définie par la somme des poids w des arêtes qui le relient avec ses voisins directs de C_i et sa compatibilité maximale noté $comp_{\max}(v)$ qui est le maximum entre la somme des poids de ses voisins libres et directs ($PVLD(v)$) et les valeurs de ses compatibilités à toutes les communautés d'une partition P :

$$comp(v, C_i) = \sum_{v' \in Ci} w(v, v') \quad (1)$$

$$comp_{\max}(v) = \max\{comp(v, C_i), \forall C_i \in P; PVLD(v)\} \quad (2)$$

Si $comp(v, C_i) = comp_{\max}(v)$ alors v est pur par rapport à C_i.

La deuxième phase est celle de l'équilibre qui vise à avoir une interaction entre les nœuds plus forte à l'intérieur d'une communauté C_i qu'à l'extérieur. Pour cela une comparaison entre la séparabilité et la compacité d'une communauté est faite :

Si $separabilité_{moy}(C_i) < compacité(C_i)$ alors la communauté C_i est équilibrée.

où $separabilité_{moy}(C_i) = moyenne\{ separabilité(C_i, C_j) > 0, \forall C_j \in P\}$ (3)

$$separabilité(C_i, C_j) = \sum_{v \in Ci,\, v' \in Cj} w(v, v') \quad (4)$$

$$compacité(C_i) = \frac{1}{2} \sum_{v, v' \in Ci} w(v, v') \quad (5)$$

Cependant, MWEP ne tient pas compte des informations dans un réseau social. En plus, dans sa première phase, on peut décider qu'un nœud v n'est pas pur par rapport à une communauté si il a un nombre élevé de voisins libres et directs ($VLD(v)$) dont la somme des poids ($PVLD(v)$) est supérieure à celle des poids de ses voisins qui se trouvent dans la communauté formée ($comp(v, C_i)$), bien qu'on peut avoir une similarité forte entre v et ses voisins de C_i. Par exemple, dans la figure 1, le nœud 5 ne sera pas ajouté à C_i bien qu'il est sémantiquement beaucoup plus proche de ses voisins de C_i que de ses voisins libres directs car $PVLD(5) > comp(5, C_i)$. En plus, dans la deuxième phase, on peut avoir à fusionner deux communautés qui ne sont pas proches sémantiquement, car la compacité d'une

communauté est comparée avec la moyenne de ses séparabilités avec les communautés qui y sont attachées.

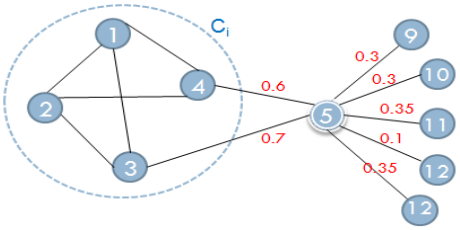


FIGURE 1 – Limite sémantique de MWEP

Par analogie avec MWEP, nous proposons une nouvelle approche SemMEP qui prend en compte plus la sémantique des informations liées aux nœuds que leur degré de connexion et cela en regroupant ensemble les nœuds très proches sémantiquement à condition qu'ils soient connectés. Dans cette approche, nous définissons la fonction de qualité «SemEP» qui dépend de deux nouvelles notions : la pureté sémantique et l'équilibre sémantique. Notre approche s'applique aux réseaux sociaux ayant des liens symétriques ne représentant pas une forte intensité structurelle, par exemple des liens d'amitié dans Facebook.

2.1 Formalisation sémantique d'un réseau social

Dans notre approche, nous représentons un réseau social comme un graphe annoté non orienté $G=(V,E,A,W)$ où V est l'ensemble de ses nœuds, E est l'ensemble de ses arêtes et A est l'ensemble des annotations des nœuds. W est une fonction qui modélise la similarité sémantique de deux nœuds qui sont obligatoirement liés. W est défini comme suit :

$$\mathrm{W}(v_i,v_j)=\mathrm{Sim}(v_i,v_j)*\delta(v_i,v_j);\forall\, v_i,v_j\in V \qquad (6)$$

où $\mathrm{Sim}(v_i,v_j)$ est la similarité sémantique des informations associées aux nœuds v_i et $\mathrm{v_j}$, $\delta(v_i,v_j)$=1 si v_i et v_j sont liés par une arête, 0 sinon.

2.2 La fonction de qualité « SemEP »

Soit une partition $P=\{C_1,...,C_k\}$ d'un réseau social représenté par $G=(V,E,A,W)$, nous définissons la fonction de qualité SemEP (Equilibre et Pureté Sémantique) comme suit :

$$SemEP(P)=\frac{1}{2}\left[SemP(P)+SemE(P)\right] \qquad (7)$$

La pureté sémantique ($SemP$) reflète l'attachement sémantique d'un membre du réseau à une communauté C_i. Pour décider si un nœud v de V est pur sémantiquement par rapport à C_i, sa similarité à C_i notée $Sim(v,C_i)$, définie par la similarité entre v et ses voisins directs de C_i, est comparée à sa similarité maximale (notée $Sim_{\max}(v)$) qui est le maximum entre sa similarité avec ses VLD, notée $\mathrm{Sim(v,C_{VLD})}$, et sa similarité avec chacune des autres communautés différentes de C_i :

$$Sim_{\max}(v)=\max\{Sim(v,C_i),\forall C_i\in P;Sim(v,C_{VLD})\} \qquad (8)$$

Si $Sim(v,C_i)=Sim_{\max}(v)$ alors v est pur sémantiquement par rapport à C_i. En cas où la similarité maximale est atteinte pour plusieurs communautés, celle qui a plus de liens avec v est favorisée. $SemP(C_i)$ est définie comme suit :

$$SemP(C_i)=\frac{|SemP(\mathrm{C_i})|}{|C_i|} \qquad (9)$$

$$avec \ \left|SemP(C_i)\right| = \left|\left\{ Sim\,(v, C_{\mathrm{i}}) = Sim_{\max}(v)\right\}\right| \qquad (10)$$

Enfin, la pureté sémantique d'un partitionnement est définie comme suit :

$$SemP(P) = \frac{1}{|P|}\sum_{C_i \in P} SemP(C_i) \qquad (11)$$

Pour vérifier si C_i est équilibrée sémantiquement, sa similarité avec chacune des autres communautés attachées à elle, noté $Sim(C_i, C_j)$ est comparée à un seuil α. Cependant, cette condition seule est insuffisante car on peut avoir deux communautés similaires mais les nœuds qui les connectent n'ont pas d'intérêt à être fusionnés, car ils sont distants sémantiquement. Il faut donc en plus comparer la similarité des deux ensembles des nœuds qui relient deux communautés (séparabilité) qui font l'objet d'une fusion, notée $Sim(C_{ij}, C_{ji})$ à un seuil β. Ci est équilibrée sémantiquement (notée $SemE(C_i)$) si $Sim(C_i, C_j) < \alpha$ ou $Sim(C_{ij}, C_{ji}) < \beta$. L'équilibre sémantique d'un partitionnement P, noté $SemE(P)$, est défini comme suit :

$$SemE(P) = \frac{\left|SemE(C_i)\right|}{|P|} \qquad (12)$$

Afin d'évaluer notre approche, nous avons adapté l'indicateur de performance CI (Connectivité Index) défini dans (Zardi & Ben Romdhane, 2013a) en proposant l'indicateur FCS (Force de Connectivité communautaire Sémantique). Il est défini avec la force de connectivité communautaire sémantique de C_i (notée $FCS_{Com}(Ci)$) qui est la force de densité sémantique de C_i. $FCS_{Com}(Ci)$ est définie en fonction de la compacité de C_i et sa séparabilité moyenne, comme suit :

$$FCS_{Com}(C_i) = \frac{Compacité(C_i) - separabilité_{moy}(C_i)}{Compacité(C_i) + separabilité_{moy}(C_i)} \qquad (13)$$

Plus la compacité de la communauté est importante et sa séparabilité moyenne est faible plus sa force de connectivité sémantique est grande quel que soit la taille de cette communauté. *FCS* d'une partition est la moyenne de $FCS_{Com}(C_i)$.

2.3 Algorithme SemMEP

SemMEP commence le processus de détection des communautés par une partition initiale dans laquelle chaque communauté contient un seul nœud. Initialement tous les nœuds sont considérés libres. Ensuite, la somme des poids connectant chaque nœud à ces voisins directs représentant un degré de poids, est calculée et les nœuds sont triés selon l'ordre décroissant de ce degré. Ces étapes représentent la phase d'initialisation de MWEP et de SemMEP. Ensuite, SemMEP comporte deux nouvelles phases pour la recherche de la pureté et de l'équilibre sémantiques, définies par analogie à la démarche algorithmique de MWEP.

Algorithme : SemMEP
Données : Un graphe Annoté et pondéré G=(V,E,A,W), une ontologie O
Résultat : Un ensemble des communautés P={C_1,…,C_k}
Début :
1. $P \leftarrow \emptyset$
2. **Pour** i=1 à |V| **faire** $C_i \leftarrow v_i$ **Fin Pour**
3. Trier les nœuds dans *Libre* selon leurs degré de poids
4. **Tant qu'ils** existent des nœuds *Libre* **faire**
   Sélectionner le premier nœud v dans Libre
   si $\exists$ C$\in$P tel que v est pur sémantiquement par rapport à C **alors**
   Ajouter v à C et Supprimer v du vecteur Libre **sinon**

Créer une nouvelle communauté C_n contenant v et ses voisins libres
Supprimer les nœuds *non purs* sémantiquement de C_n
Ajouter à C_n les voisins purs *de chaque nouveau membre et* Ajouter C_n àP
Supprimer les membres de C_n du vecteur Libre
Fin Tant que

5. **Tant qu'**il existe des communautés *non équilibrées sémantiquement* **faire**
Fusionner chaque communauté non *équilibrée* avec la plus similaire
Déplacer les sommets qui deviennent impurs vers les groupes dans lesquels ils seront purs sémantiquement
Fin Tant que

Fin.

3 Expérimentation

Pour expérimenter notre méthode, nous avons considéré le réseau « Karaté » que nous avons annoté avec une ontologie qui définit 4 centres d'intérêt (Cinéma, Sport, Informatique et Musique), selon 4 scénarios d'annotation (S1, S2, S3, S4). Nous avons appliqué des mesures de similarité sémantique adéquates sur cette ontologie, pour mesurer la similarité de deux concepts ou de deux groupes de concepts qui annotent des nœuds ou des communautés (Blanchard et al. 2008) (Harzallah et Berio, 2015). Nous avons ensuite appliqué notre approche et l'approche sémantique de Dang et Viennet (2012) (notée dans la suite ADV) pour la détection des communautés sur ces scénarios. ADV utilise l'algorithme de Louvain pour optimiser la fonction de qualité « Modularité » notée Q, et la densité et l'entropie comme indicateurs de performance. La table1 illustre les résultats de cette expérimentation.

TABLE 1 – *Résultats de SemMEP et de ADV*

Approche / Mesures	ADV				SemMEP			
	S1	**S2**	**S3**	**S4**	**S1**	**S2**	**S3**	**S4**
NC	3	4	4	3	2	4	3	4
Q/SemEP	0.53	0.58	0.58	0.55	1	1	1	1
Densité	0.80	0.73	0.73	0.80	0.87	0.73	0.78	0.71
Entropie	0	0.01	0.01	0.37	0	0	0.07	0.29
FCS	0.68	0.61	0.61	0.78	1	0.70	0.78	0.60

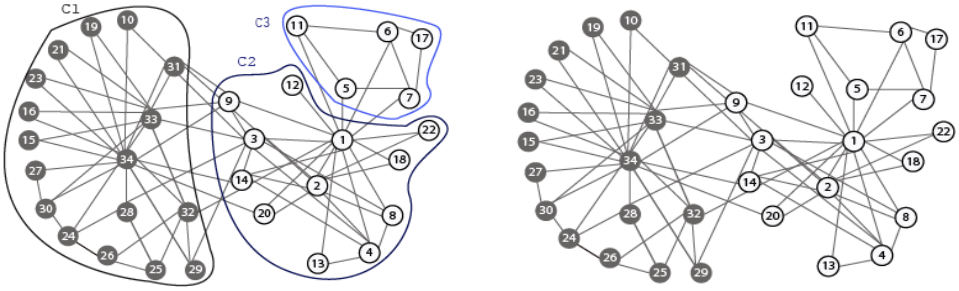


FIGURE 2 – (a) Partition de S1 avec ADV (b) Partition de S1 avec SemMEP

Dans S1, le réseau est annoté par deux centres d'intérêt : Musique (nœud grisé) et Sport (nœud transparent) dans la figure 2. Pour S1, SemMEP a identifié 2 communautés, la première s'intéresse à la musique et la deuxième au sport (figure 2- b), ADV a déterminé 3 communautés illustrées par (figure 2- a). SemEP est optimale alors que Q= 0.53. Les valeurs de la densité et de FCS sont plus élevées pour SemEP que pour ADV. Dans S2, nous avons utilisé le résultat de la figure (2-a) et nous avons annoté les membres associés aux nœuds : 25, 26, 28 et 32 de C1 par le concept « Jazz » qui est une spécialisation de « musique » et les

nœuds associés aux membres de C3 par « cyclisme » qui est une spécialisation de « sport ». L'annotation des autres nœuds n'a pas changé par rapport à S1. Comme attendu, SemMEP a déterminé une partition avec 4 communautés : les nœuds 25, 26, 28 et 32 de C1, C2, C3 et les autres nœuds de C1. ADV a déterminé presque les mêmes communautés avec la différence que le nœud 29 annoté par la musique a migré vers la communauté qui s'intéresse au Jazz, ce qui a rendu cette communauté non strictement homogène. S3 est identique à S2 mais les seuils de similarités $\alpha et \beta$ ont été baissés de 1 à 0.5. Dans ce cas SemMEP a déterminé 3 communautés en fusionnant celles qui s'intéressent au sport et au cyclisme. S4 est une annotation aléatoire du réseau. Pour ces différents scénarios, les résultats de SemMEP pour les différentes mesures sont meilleurs que ceux de ADV, sauf pour la densité (S4). Tous ces résultats montrent que notre modèle SemMEP est plus intéressant que celui de ADV : il a donné un partitionnement plus homogène sémantiquement que celui de ADV.

4 Conclusion

Nous avons proposé une nouvelle approche SemMEP pour la détection des communautés dans un réseau social qui prend en compte plus les aspects sémantiques dans un réseau social que ses aspects structurels en regroupant ensemble les acteurs qui sont proches sémantiquement. Notre approche évalue la sémantique des informations dans un réseau avec une ontologie et des mesures sémantiques appliquées aux concepts de cette ontologie qui annotent les nœuds du réseau. En plus, elle a intégré ces aspects sémantiques dans différentes étapes de processus de détection. Des expérimentations préliminaires ont montré que notre approche donne des résultats plus intéressants que ceux de ADV. Dans nos travaux futurs, nous allons la valider sur des graphes à grand échelle.

Références

BLANCHARD E, HARZALLAH M et KUNTZ P. (2008). A generic framework for comparing semantic similarities on a subsumption hierarchy. *18th European Conference on Artificial Intelligence (ECAI)*, pp 20-24.

BLONDEL V. D., GUILLAUME J.-L., LAMBIOTTE R. et LEFEBVRE E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no 10, p. P10008 (12pp).

CRUZ J. D., BOTHOREL C. et POULET F. (2013). Community detection and visualization in social network: integrating structural and semantic information. *ACM transaction on intelligent systems and technology (TIST)*, vol. 5, n°1, pp. n°11.

DANG T. & VIENNET E. (2012). Community detection based on structural and attribute similarities. *In International conference on digital society (icds)*, p. 7–14.

ERETEO G., GANDON F. et BUFFA M. (2011). SemTagP: Semantic Community Detection in Folksonomies. *IEEE/WIC/ACM International Conference on Web Intelligence*, Lyon.

HARZALLAH M. & BERIO G.(2015). A Unified Framework for Semantic Comparison of Objects: Extension to Semantic Graphcomparison. *KES 2015*: 547-556

LEPROVOST D, ABROUK L. et GROSS-AMBLARD D. (2012). Discovering implicit communities in web forums through ontologies. *Web Intelligence and Agent Systems*, 10(1) :93–103.

WAN T, WEI LIU W& ZONGTIAN LIU, LIU Z. (2014). A community discovering method based on event network for topic detection. *Advanced Communication Technology (ICACT)*.

ZARDI H. & BEN ROMDHANE L. (2013a). An O(n2) algorithm for detecting communities of unbalanced sizes inlarge scale social networks. *Knowl.-Based Syst.* 37: 19-36

ZARDI H. & BEN ROMDHANE L. (2013b). WMEP: Efficiently Mining Community Structures in WeightedLarge Scale Social Graphs. *In The first International Conference on Reasoning and Optimization in Information Systems (ROIS'2013),* Hammam-Sousse, Tunisia, Sept 6–7, 2013.

ZHANG F., LI J., LI F, XU M, XU R. et HE X. (2015). Community Detection Based on Links and Node Features in Social Networks. *Lecture Notes in Computer Science*, pp 418-429.

Vers une Ingénierie des Connaissances Personnelles – Étude de cas pour l'organisation des collections musicales

Nicolas Greffard, Pascale Kuntz, Éric Languénou

LINA/UNIVERSITÉ DE NANTES
2, rue de la Houssinière, BP 92208
44322 Nantes Cedex 03, France
[nicolas.greffard, pascale.kuntz,eric.languenou]@univ-nantes.fr

Résumé : Le traitement de nos informations numériques personnelles est devenu une tâche majeure de nos vies contemporaines et la complexité des informations stockées nécessite des approches permettant de structurer et de gérer les connaissances associées à ces dernières. Dans cette communication, nous nous focalisons sur une étape amont préalable à la proposition d'outils d'assistance personnalisés : l'identification des processus d'organisation (ici de classement) dans un environnement familier. Le terrain d'expérimentation est celui des collections musicales personnelles stockées sur les disques durs de nos ordinateurs. Une analyse de la structuration de ces collections nous a permis de mettre en évidence un ensemble restreint de stratégies d'organisation qui affinent celles étudiées dans la littérature pour d'autres types de collections. Notre discussion ouvre sur les nouveaux défis associés à l'émergence du syndrome du « big data » à l'échelle individuelle.

Mots-clés : données personnelles, classification, collections musicales.

1 Introduction

Les terrains privilégiés par l'ingénierie des connaissances (IC) sont les entreprises et les administrations (Charlet, 2005). Ce positionnement s'explique à la fois par l'historique des systèmes experts et des systèmes à base de connaissances mais aussi par les besoins actuels des institutions qui ne cessent de croître. Pourtant la définition de l'IC n'est pas si restrictive dans sa cible et peut s'appliquer aujourd'hui, nous semble-t-il, au champ individuel. En effet, le traitement de nos informations numériques personnelles est devenu une tâche majeure de nos vies contemporaines et la volumétrie et la complexité des informations stockées nécessitent des méthodes et techniques permettant de structurer et de gérer les nouvelles connaissances associées à ces informations ; c'est-à-dire une ingénierie des connaissances adaptée à l'échelle personnelle. Si nous n'avons pas trouvé trace de ces préoccupations dans les actes de la conférence IC de cette dernière décennie, des questionnements sur l'impact du virage numérique dans le traitement des connaissances individuelles sont au cœur des débats de la communauté structurée autour du « Personal Information Management » (PIM) (Jones, 2008) dont l'attention a porté initialement principalement sur les documents, les courriels et les pages Web. L'impact du numérique sur l'accès et le stockage dans la sphère privée stimule aujourd'hui les recherches sur l'organisation des collections personnelles (Lee, 2011) associées notamment aux bibliothèques numériques, aux photothèques et vidéothèques personnelles. Dans cet article, nous nous penchons sur un champ qui a été très peu étudié : celui de l'organisation des collections musicales personnelles. Cette absence relative contraste avec l'importance de la musique dans nos vies quotidiennes (DeNora, 2000). Aujourd'hui l'attention médiatique porte majoritairement sur le

streaming car ses revenus sont en croissance mais, pourtant, le téléchargement représente encore 49% du marché et une enquête récente de l'Hadopi confirme l'attachement individuel des jeunes à la constitution des bibliothèques numériques de biens culturels. Dans ce contexte, nos travaux se focalisent ici sur l'organisation des discothèques numériques personnelles.

Plus précisément, l'objectif de cet article est d'identifier les modes d'organisation mis en œuvre dans le classement des collections musicales stockées sur nos disques personnels. Cette identification des comportements développés dans un environnement familier nous semble un préalable nécessaire à la proposition d'outils d'assistance personnalisés. Après un état de l'art synthétique sur les travaux – assez rares – consacrés à ces collections, nous présentons une étude effectuée sur un échantillon de disques durs recueillis essentiellement chez des jeunes adultes et nous discutons des perspectives nouvelles que cette démarche ouvre en ingéniérie des connaissances.

2 État de l'art

De nombreux outils de gestion des données musicales ont été développés dans la communauté MIR (Music Information Retrieval). Ces travaux se focalisent essentiellement sur la classification automatique des données et sur la visualisation des classes produites ou des proximités calculées (Li *et al.*, 2012). Mais, comme le souligne un article récent au titre explicite « The neglected user » (Schedl *et al.*, 2013), l'utilisateur est peu pris en compte dans la démarche. Il ne l'est souvent qu'*a posteriori* pour les tests de validation d'interfaces et son degré de liberté dans la gestion des données est fortement contraint par les topologies des espaces de classement définis dans les outils et les modalités d'interaction. De plus, comme le souligne Jones (Jones, 2008) il faut distinguer l'organisation de la gestion de l'information et la majorité des travaux porte plus sur la question de la gestion des données pilotée par des interfaces élaborées que sur celle de l'organisation personnelle sur des supports familiers.

En fait, à notre connaissance, la composition des discothèques numériques personnelles a été peu abordée dans la littérature (Sease & McDonald, 2009; Brinegar & Capra, 2010; Kamalzadeh *et al.*, 2012; Lee & Waterman, 2012; Jacques, 2015). Les données sur lesquelles ces travaux reposent sont issues d'enquêtes par questionnaires complétées souvent par quelques entretiens ethnographiques. Les observations confirment la difficulté d'estimation de la taille des collections à l'échelle individuelle et la variabilité des situations observées. Si ces enquêtes permettent de commencer à défricher les pratiques, elles se heurtent cependant aujourd'hui à deux écueils. La quantité d'informations auxquelles les natifs du numérique sont aujourd'hui quotidiennement exposés rend leur description verbale de plus en plus délicate. Toutes proportions gardées, on se retrouve face à un phénomène de type « big data » à l'échelle individuelle. De plus, avec les nouvelles modalités de « l'exposition de soi » sur les réseaux sociaux, il est légitime de s'interroger sur la véracité des contenus exprimés lors d'une enquête. Pour palier à ces difficultés, nous avons donc choisi ici d'observer *in silico* les contenus des disques durs des ordinateurs utilisés au quotidien sur lesquels on trouve des fichiers musicaux.

3 La méthodologie

Les données étudiées ont été recueillies auprès de 32 jeunes individus, pour la majorité étudiants, âgés de 17 à 30 ans, volontaires pour que le disque dur de leur ordinateur personnel principal (portable ou fixe) soit scanné pour le recueil des informations. Un outil logiciel parcourt un disque en enregistrant l'emplacement, le nom et la date de création de tous les fichiers musicaux pour les principaux formats et filtre les fichiers musicaux de petites tailles (< 1Mb) relatifs au système d'exploitation. Nous avons ainsi recueilli 10 4171 fichiers musicaux, soit une moyenne de 3255 morceaux par participant. Cependant, la moyenne est peu informative de par l'hétérogénéité de la distribution du nombre de fichiers/participants qui suit une loi proche d'une loi de Pareto.

Influencés par de nombreux travaux sur l'analyse de réseaux où la caractérisation de la topologie des relations tente de servir de révélateur à une organisation cachée plus subtile, nous avons commencé par étudier les arborescences des fichiers associés à chaque collection. Ces dernières se reconstruisent facilement à partir des données recueillies. Nous avons effectué une classification hiérarchique – non supervisée – de ces arborescences. Puis, nous avons tenté d'analyser le contenu – selon les genres musicaux – des différentes classes extraites de la classification et d'identifier les stratégies personnelles d'étiquetage des répertoires.

Analyse structurelle. Afin de découvrir les différentes stratégies mises en œuvre dans l'organisation des discothèques, nous avons effectué une classification des arbres à partir d'une description de ces derniers par des indicateurs combinatoires classiques : degré moyen des noeuds μ_d et écart-type associé, profondeur moyenne des nœuds μ_p et écart-type, longueur moyenne μ_l des chaînes entre la racine et les feuilles et écart-type, nombre de feuilles n_f (i.e nombre de fichiers musicaux) et nombre de nœuds n_r (i.e. nombre de répertoires). La dissimilarité entre deux collections est mesurée par la distance euclidienne entre les valeurs des indicateurs normalisés. Nous avions préalablement considéré une distance de Kullback-Leibler entre les distributions des différents indicateurs sans que cela n'apporte de précision supplémentaire dans les résultats obtenus. La classification a été effectuée par une méthode hiérarchique classique basée sur le critère de Ward.

Analyse "sémantique". Ayant récolté près de 100 000 noms de fichiers musicaux, une préclassification en un nombre de classes plus raisonnable nous est apparue indispensable pour une interprétation humaine. Un grand nombre des travaux de sociologie consacrés aux pratiques de l'écoute musicale dont nous avons eu connaissance font référence aux « goûts musicaux » et ces derniers sont définis par des « genres musicaux » . Dans les enquêtes par questionnaire classiques telles que celles menées par le Ministère de la Culture, une liste fermée de genres pré-définis est proposée. Mais, dans les fichiers numériques, le genre n'est pas donné ou est très mal renseigné et il nous a fallu recourir à une étape de classification. Malgré le développement de nombreux modèles computationnels, le problème de la détermination automatique du genre d'un morceau musical reste un problème délicat largement ouvert (Seyerlehner *et al.*, 2010), et à notre connaissance, les musicologues n'ont pas produit une ontologie consensuelle des genres musicaux qui pourrait servir de support à un codage automatique. Ce problème sortant largement du cadre de ce travail, nous avons choisi, pour des raisons opérationnelles, de recourir à la

base de données d'EchoNest [1] qui propose un étiquetage automatique des artistes en genres (sur une base de plus de 700 genres). L'extraction des noms d'artistes de nos données a été facilitée par les méta-données et nous avons ainsi recueilli 5405 noms d'artistes différents. L'étiquetage d'EchoNest repose principalement sur une analyse fréquentielle des termes provenant de publications en rapport avec les artistes collectées principalement sur des sites spécialisés et des réseaux sociaux. À chaque artiste apparaissant dans les métadonnées, nous avons associé via la consultation d'EchoNest le genre indiqué et nous avons conservé *in fine* pour chaque artiste le genre majoritairement associé sur l'ensemble des collections ; à chaque artiste est donc associé un seul genre musical. Sur notre échantillon, 478 genres musicaux sont représentés avec une moyenne de 67 genres par discothèque et une forte disparité.

Etiquetage manuel des répertoires dans les discothèques personnelles. En sus de la classification automatique des contenus en genres musicaux, nous avons tenté de mieux comprendre la procédure manuelle de classement en analysant les noms donnés aux répertoires. La très forte variabilité à la fois des noms proposés et de leurs orthographes a rendu la tâche difficile. En nous basant sur les résultats des enquêtes citées dans l'état de l'art, nous nous sommes focalisés sur trois classes d'étiquettes : album, artiste et genre. L'identification des albums a été effectuée en deux étapes : un filtrage manuel basé sur nos connaissances et un filtrage automatique basé sur un étiquetage automatique classiquement utilisé sur les sites de téléchargement « artiste-album-titre » et la position du répertoire dans la hiérarchie. Bien que des erreurs puissent persister, les différences entre les résultats sont suffisamment importantes pour ne pas altérer l'interprétation.

4 Classification des collections

La classification permet de faire apparaître trois classes distinctes correspondantes à des stratégies d'organisation des discothèques personnelles différentes :

— La classe 1, la plus peuplée, regroupe des discothèques basées sur un classement dichotomique qui combine des fichiers bien rangés et des gros répertoires « fourre-tout » dont le volume est supérieur à l'ensemble des autres catégories et contiennent deux fois plus d'artistes et de genres que tous les autres répertoires des discothèques de l'échantillon. De plus, dans ces discothèques les genres semblent plus spécialisés. Le nombre médian m_g de morceaux musicaux par genre dans toutes les discothèques est de 8 alors que dans la classe 1 seuls 24% des genres comportent plus de m_g morceaux. Un test de Wilcoxon confirme que cette différence est statistiquement significative.
— La classe 2 regroupe des discothèques de "power users" qui conservent un grand nombre de fichiers musicaux stockés dans une arborescence très structurée avec de très nombreux répertoires organisés eux-mêmes de façon récursive en sous-répertoires. Les discothèques de cette classe sont celles qui contiennent le plus grand nombre de fichiers musicaux (8408 en moyenne) et elles sont également les plus hétéroclites en terme de goûts puisque cette classe comporte 396 genres différents (40% de plus que la classe 1 et deux fois plus que la classe 3) et ces genres sont très fournis : 185 sont associés à plus de m_g fichiers avec une différence significative avec les autres classes.

1. ww.echonest.com

— La classe 3 se différencie des autres principalement par la faible profondeur de ses répertoires. Une analyse complémentaire des dates de création de ces derniers montre qu'en moyenne 56% de ces répertoires contiennent des fichiers ajoutés au même moment — contre 27% pour les autres classes—. Il s'agit donc d'un comportement de classement qui suit la stratégie « je range tout de suite en créant une nouvelle classe à chaque fois ». Et les discothèques de cette classe sont celles qui comportent le moins de genres différents. Une analyse des genres les plus représentés relève des genres qui sortent plus des courants dominants.

Cette typologie des comportements de classement a été complétée par l'analyse des noms donnés aux répertoires. L'album joue un rôle majeur dans le processus de classement ; ce qui peut différer des données verbales recueillies dans les enquêtes par questionnaire (Kamalzadeh *et al.*, 2012) où le genre a un rôle souvent beaucoup plus important. La typologie classique en trois items (genre, artiste, album) ne concerne que 80% des étiquettes et on note l'émergence de noms liés aux actions liés à la musique (« pour danser », « pour courir »). Cette pratique rappelle celle qui avait été déjà été mise en évidence avec des documents physiques où les attributs du classement ne dépendaient pas uniquement du contenu « objectif » des documents (e.g. auteur, sujet, titre) mais de l'interaction entre l'utilisateur et l'information (Kwasnik, 1991).

Cette première analyse permet de révéler des stratégies d'organisation différentes. Dans la littérature, qu'il s'agisse de classements d'artéfacts physiques ou numériques, bon nombre de travaux se réfèrent à la distinction introduite dans les travaux pionniers de Malone (Malone, 1983) sur l'organisation des documents au bureau entre les dossiers (« files »), où les documents sont étiquetés et ordonnés, et les piles (« piles ») où les documents sont posés sans ordre spécifique. Quel que soit le type d'information, la création de piles demande un effort minimal alors que la création de dossiers requiert un effort cognitif et manuel plus important (Jones, 2008). Nos résultats recouvrent la distinction classique (« neat » et « messy ») et l'affinent. Au niveau individuel une organisation méticuleuse peut se combiner avec une organisation négligée. Au niveau de l'échantillon, si des différences persistent, la combinatoire des modes d'organisation est finalement assez restreinte ; ce qui rend la perspective d'une assistance personnalisée plus aisée.

5 Conclusion

Dans cet article, nous avons proposé une analyse descriptive de l'organisation des collections musicales à partir de l'observation d'empreintes numériques (ici les arborescences de fichiers musicaux stockées sur les disques durs personnels). Il s'agit d'une première étape visant à identifier les processus mis en œuvre dans un environnement numérique familier. Nous avons mis en évidence différents modes de classement dont la robustesse devra être testée sur d'autres échantillons. Mais les facteurs explicatifs restent largement à explorer. La ré-utilisabilité reste une motivation souvent énoncée dans la littérature mais elle n'est pas la seule (Kaye *et al.*, 2006). L'identification des liens ou de l'absence de lien entre les modes de classement et les objectifs de l'utilisateur reste évidemment une question essentielle pour le développement d'une Ingénierie des Connaissances Personnelles.

Pour notre étude, nous nous sommes restreints aux données stockées sur un type de support spécifique mais aujourd'hui le nombre de supports qu'un individu a à sa disposition ne

cesse de croître, et s'ajoute aux dispositifs personnels, le stockage distribué sur des supports externes. Une communication récente de S. Abiteboul (détaillée dans Abiteboul *et al.* (2015)) au titre évocateur « Quand nos vies numériques deviennent des bases des connaissances » ouvre des nouveaux défis passionnants pour l'Ingénierie des Connaissances. La construction de ces ontologies personnelles revisite en profondeur la démarche classique dont la légitimité repose souvent sur un consensus entre experts d'un domaine. Et elle renouvelle l'intérêt d'une construction automatique qui reste encore aujourd'hui bien délicate.

Références

ABITEBOUL S., ANDRÉ B. & KAPLAN D. (2015). Managing your digital life. *Commun. ACM*, **58**(5), 32–35.

BRINEGAR J. & CAPRA R. (2010). Understanding personal digital music collections. *Proc. of the American Society for Information Science and Technology*, **47**(1), 1–2.

CHARLET J. (2005). L'ingénierie des connaissances, entre science de l'information et science de gestion. In *Entre la connaissance et l'organisation, l'activité collective*, p. 306–309. P. Lorino & R. Teulier Eds, La Découverte.

DENORA T. (2000). *Music in Everyday Life*. Cambridge University Press.

JACQUES J. (2015). Les pratiques d'organisation des collections musicales numériques comme enjeu central de l'écoute contemporaine. In *Colloque Musimorphose, Paris (actes à paraître)*.

JONES W. (2008). *Keeping Found Things Found : The Study and Practice of Personal Information Management : The Study and Practice of Personal Information Management*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

KAMALZADEH M., BAUR D. & MÖLLER T. (2012). A survey on music listening and management behaviours. In *Proc. of the 13th Int. Symp. on Music Information Retrieval*, p. 373–378.

KAYE J. J., VERTESI J., AVERY S., DAFOE A., DAVID S., ONAGA L., ROSERO I. & PINCH T. (2006). To have and to hold : Exploring the personal archive. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, p. 275–284 : ACM.

KWASNIK B. H. (1991). The importance of factors that are not document attributes in the organization of personal documents. *The School of Information Studies Faculty Scholarship*.

LEE C. (2011). *I, Digital : Personal Collections in the Digital Era*. The Society of American Archivist.

LEE J. H. & WATERMAN N. M. (2012). Understanding user requirements for music information services. In *Proc. of the 12th Int. Symp. on Music Information Retreival*, p. 253–258.

LI T., OGIHARA M. & TZANETAKIS G. (2012). *Music Data Mining*. Chapman & Hall/CRC.

MALONE T. W. (1983). How do people organize their desks ? implications for the design of office information systems. *ACM Trans. Inf. Syst.*, **1**(1), 99–112.

SCHEDL M., FLEXER A. & URBANO J. (2013). The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, **41**(3), 523–539.

SEASE R. & MCDONALD D. W. (2009). Musical fingerprints : collaboration around home media collections. In *Proc. of the ACM international conference on Supporting group work*, p. 331–340.

SEYERLEHNER K., WIDMER G. & KNEES P. (2010). A comparison of human, automatic and collaborative music genre classification and user centric evaluation of genre classification systems. In *Adaptive Multimedia Retrieval. Context, Exploration, and Fusion, 2010*, p. 118–131.

Démonstrations et Posters

Du TALN au LOD : Extraction d'entités, liage, et visualisation

Cédric Lopez[1], Osmuk[1], Dana Popovici[1], Farhad Nooralahzadeh[2], Domoina Rabarijaona[1], Fabien Gandon[2], Elena Cabrio[2], Frédérique Segond[1]

[1] VISEO TECHNOLOGIES, R&D, Grenoble, France
firstname.name@viseo.com

[2] INRIA, Wimmics, Sophia Antipolis, Nice, France
firstname.name@inria.fr

Résumé : Dans un contexte de veille stratégique, nous avons développé un prototype prenant la forme d'un plugin de navigateur ayant pour principale ambition d'enrichir les connaissances des utilisateurs naviguant sur le Web. Au fur et à mesure de la navigation sur le Web, le système peuple la base de connaissance et tisse des liens avec le Web des données ouvertes que l'utilisateur peut parcourir. Ce prototype s'appuie et démontre en pratique des techniques d'extraction d'entités d'intérêt et de leurs relations dans une page Web couplées à une représentation des connaissances extraites au format du web sémantique et liées avec des données du Linked Open Data. Finalement, le plugin propose une visualisation en temps réel de l'ensemble de ces données liées en regard des pages consultées.

Mots-clés : Extraction d'entités nommmées, Données ouvertes, Ontologies, Visualisation.

Un des objectifs du laboratoire commun SMILK (Social Media Intelligence and Linked Knowledge, LabCom ANR) concerne l'étude du couplage du Traitement Automatique du Langage Naturel (TALN) au Linked Open Data (LOD). Pour atteindre cet objectif, nos recherches portent sur : 1) l'extraction d'entités d'intérêt et de leurs relations dans un contenu textuel non structuré, 2) la représentation des connaissances extraites, 3) le liage des données extraites avec les données du LOD, 4) la visualisation et l'exploration des données liées.

Pour valider nos recherches et démontrer les possibilités nous avons développé un prototype qui prend la forme d'un plugin de navigateur ayant pour ambition d'enrichir les connaissances des utilisateurs naviguant sur le Web. Dans un contexte de veille stratégique, notre cas d'application se concentre sur le secteur de la cosmétique, bien représenté chez Viseo par des clients d'envergure tels que L'Oréal, L'Occitane, ou LVMH (Moët Hennessy Louis Vuitton).

Le prototype SMILK analyse les pages Web en vue d'identifier des entités pertinentes dans le domaine de la cosmétique. Celles-ci correspondent à des classes de notre ontologie ProVoc (*PROduct VOCabulary*) qui a pour vocation la publication de données liées portant sur des produits sur le Web (Lopez *et al.*, 2016), précisément : les groupes (ex : *L'Oréal*), leurs divisions (ex : *L'Oréal produits grand public*), les marques (ex : *L'Oréal Paris*), les gammes de produits (ex : *Elsève*) et les noms et caractéristiques de produits (ex : *Color Vive 200*).

La reconnaissance automatique des entités d'intérêt et de leurs relations est effectuée par Renco, notre système à base de règles linguistiques (Lopez *et al.*, 2014). Les règles développées sont de type lexico-syntaxique, fondées sur les principes de (McDonald, 1996) et (Hearst, 1992) qui considèrent les contextes droit, gauche, et interne de l'entité pour la désambiguïser. Un résultat de cette étape de reconnaissance d'entités est illustré à la Figure 1.

Les entités repérées sont ensuite liées au LOD (précisément DBpedia) par notre système de liage des données fondé sur le framework Dexter (Ceccarelli *et al.*, 2013), pour lequel nous

FIGURE 1 – Un résultat de l'extraction des entités d'intérêt sur une page Web de Vogue.fr.

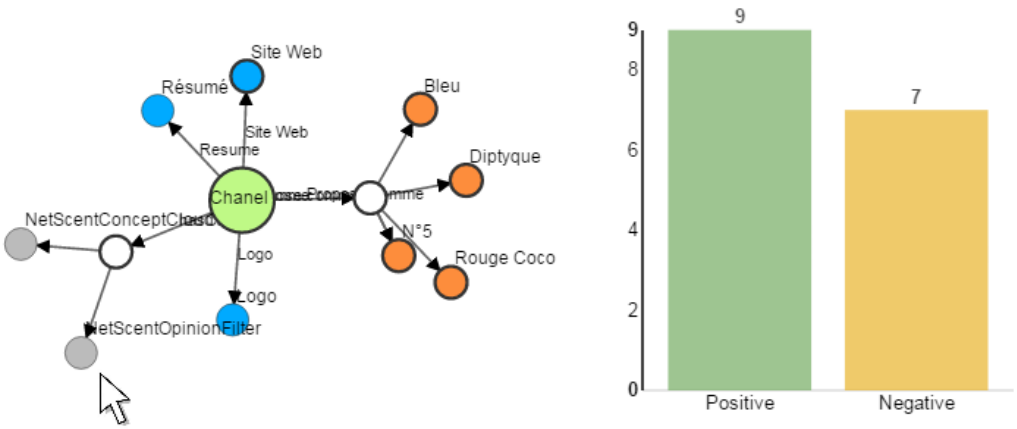


FIGURE 2 – Liage des entités extraites avec les données de bases de connaissance privées et publiques. En bleu, données provenant de DBpedia ; en gris : données provenant de NetSent ; en orange : données de notre base de connaissance privée

avons retenu notre approche de propagation sémantique (Nooralahzadeh *et al.*, 2016). Le système peuple ainsi semi-automatiquement notre base de connaissance au format RDF au fur et à mesure de la consultation des pages Web. Une validation manuelle peut être effectuée par l'administrateur de la base de connaissance afin d'assurer que les données qui y sont intégrées sont correctes.

Un exemple de graphe RDF, généré à la volée par le plugin, est montré en Figure 2. Au cours de la navigation dans le graphe, plusieurs liens apparaissent : vers notre base de connaissance privée, vers DBpedia, ou vers NetSent, une base de connaissance fournissant des résultats d'analyse d'opinion réalisée sur des forums de cosmétiques et développée en collaboration avec Holmes Semantic Solutions [1].

Par ailleurs, notre base de connaissance peut être explorée en utilisant notre outil SMILK Viewer qui repose sur le serveur RDF Jena Fuseki. L'accès au graphe de connaissance s'opère en choisissant une entité particulière dans la liste des entités catégorisées par type (groupes, divisions, marques, etc.). L'utilisateur peut ensuite facilement y naviguer et découvrir les connaissances recueillies par le plugin au fil des pages Web parcourues, ainsi que des données DBpedia et Netscent s'y rapportant. Par exemple, en sélectionnant la marque *Chanel*, apparaissent différents produits de ladite marque dont *N° 5* de type *eau de parfum*. En cliquant sur *N° 5*, le graphe permet de prendre connaissance de 3 de ses ambassadeurs dont *Audrey Tautou* pour laquelle le système nous informe qu'elle est une actrice française née le 9 août 1976 à Beaumont.

Dans cette démonstration nous naviguerons sur des pages Web évoquant des produits pour montrer comment le plugin reconnait les entités qui y sont mentionnées et collecte des données supplémentaires. Nous verrons ainsi les traitements effectués sur les textes et parcourrons le graphe construit et visualisé en regard de pages, démontrant l'apport de ces sources extérieures à l'augmentation de la compréhension des pages visitées.

Remerciements. Ce travail est réalisé dans le cadre du Laboratoire Commun SMILK financé par l'ANR (ANR-13-LAB2-0001).

Références

CECCARELLI D., LUCCHESE C., ORLANDO S., PEREGO R. & TRANI S. (2013). Dexter : an open source framework for entity linking. In *Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval*, p. 17–20 : ACM.

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, p. 539–545 : Association for Computational Linguistics.

LOPEZ C., NOORALAHZADEH F., CABRIO E., SEGOND F. & GANDON F. (2016). Provoc : une ontologie pour décrire des produits sur le web. In *Actes d'IC'16*, p. to appear.

LOPEZ C., SEGOND F., HONDERMARCK O., CURTONI P. & DINI L. (2014). Generating a resource for products and brandnames recognition. application to the cosmetic domain. In *Actes d'LREC'14*, p. 2559–2564.

MCDONALD D. (1996). Internal and external evidence in the identification and semantic categorization of proper names. *Corpus processing for lexical acquisition*, p. 21–39.

NOORALAHZADEH F., LOPEZ C., CABRIO E., GANDON F. & SEGOND F. (2016). Adapting semantic spreading activation to entity discovery in text. In *Actes d'NLDB'16*, p. to appear.

1. http ://www.ho2s.com/fr/

PERSOREC : un système personnalisé de recommandations pour les folksonomies basé sur les concepts quadratiques

Mohamed Nader Jelassi[1,2,3], Sadok Ben Yahia[1] et Engelbert Mephu Nguifo[2,3]

[1] Université Tunis El Manar. Faculté des Sciences de Tunis, Tunis, Tunisie.

[2] Clermont Université, Université Blaise Pascal, LIMOS, BP 10448, F-63000 Clermont-Ferrand, France.

[3] CNRS, UMR 6158, LIMOS, F-63171 Aubière, France.
{nader.jelassi@isima.fr, sadok.benyahia@fst.rnu.tn, engelbert.mephu_nguifo@univ-bpclermont.fr}

Résumé : Nous proposons un nouveau système appelé PersoRec afin de personnaliser les recommandations (d'amis, de tags ou de ressources) faites aux utilisateurs dans les folksonomies. La personnalisation des recommandations est réalisée en prenant en compte le profil des utilisateurs. PersoRec est donc capable de générer une recommandation personnalisée pour chaque utilisateur selon le mode de recommandation qu'il désire (recommandation d'amis, de tags ou de ressources) et selon le profil qu'il possède.

Mots-clés : folksonomie, personnalistion, recommandation, concepts quadratiques, profil.

1 Introduction et Motivations

Une *folksonomie* désigne un système de classification collaborative par les internautes (Mika, 2007). L'idée est de permettre à des utilisateurs de partager et de décrire des objets via des mots-clés (tags) librement choisis. Les *folksonomies* ont à tenir compte des besoins de ses utilisateurs lors de la recommandation de tags ou de ressources. Cela a incité les chercheurs à proposer des systèmes de recommandation personnalisés. En effet, le domaine de personnalisation tente de fournir des solutions afin d'aider les utilisateurs à partager les bons tags et les bonnes ressources parmi le très grand nombre de données dans les *folksonomies* Qumsiyeh & Ng (2012) Kim *et al.* (2011) Bellogín *et al.* (2013). De plus, la personnalisation tente d'aider les utilisateurs à aborder le problème de surcharge d'information Das *et al.* (2012). Pour atteindre cet objectif, nous considérons une nouvelle dimension dans une *folksonomie* qui contient des informations supplémentaires sur les utilisateurs, classiquement composée de trois dimensions (utilisateurs, tags et ressources), et nous proposons une approche de regroupement des utilisateurs aux intérêts équivalents sous forme de structure appelées concepts quadratiques (Jelassi *et al.*, 2013). Un concept quadratique illustre une conceptualisation partagée dans la *folksonomie* contenant un ensemble d'utilisateurs ayant partagé en commun un certain nombre de tags et de ressources et ayant le même profil. Un exemple de quadri-concept serait : *"Jack et Kate qui sont âgés entre* 18 *et* 25 *ans utilisent les tags 'action' et 'aventure', parmi d'autres, pour annoter des films comme 'Indiana Jones' et 'Star Wars'"*. Une fois extraits, ces quadri-concepts sont utilisés pour notre algorithme de recommandation personnalisée. De plus, les utilisateurs des *folksonomies* sont intéressés par des recommandations multi-mode (utilisateurs, tags et ressources), ainsi, les algorithmes répondant à cette attente sont appréciés (Ricci *et al.*, 2011). Enfin, nous nous intéressons à une minorité de nouveaux utilisateurs qui risquent de ne pas recevoir de re-

commandations. Cette problématique est connue comme le "démarrage à froid" (ou cold start) (Ricci *et al.*, 2011).

2 PERSOREC : un système personnalisé de recommandation basé sur les quadri-concepts

PERSOREC prend un ensemble de quadri-concepts $\mathcal{QC}$ comme entrée ainsi qu'un utilisateur cible u avec son profile v et (optionellement) une ressource r (à annoter). PERSOREC permet de générer trois ensembles : un ensemble d'utilisateurs proposés, un ensemble de tags suggérés et un ensemble de ressources recommandées. En plus de l'algorithme original, nous avons ajouté une mesure de score afin de classifier les recommandations par ordre d'importance. La mesure (notée rec_score) correspondant à un profil v est définie comme suit :

$$rec\_score(r_i, v) = \frac{|u_i|}{|UU|} / \exists t_i \exists r_i \exists v_i \in v, (u_i, t_i, r_i, v_i) \in \mathcal{F}_v \quad (1)$$

La mesure rec_score d'une ressource r_i correspondant à un profil v est le nombre d'utilisateurs uniques ayant le même profil v (ou au moins une information de profil $v_i \in v$) et ayant partagé la même ressource r_i, divisé par le nombre total d'utilisateurs uniques dans l'ensemble des quadri-concepts (noté UU). Par exemple, si une ressource r_1 a été partagée par 7 utilisateurs différents parmi un ensemble total de 67 utilisateurs uniques, son score sera égal à 0.104.

FIGURE 1 – Un instantané du siteweb PERSOREC pour le jeu de données MOVIELENS. (**gauche**) le profil de *Yasmine*, ses films partagés et sa liste d'amis (**centre**) les recommandations de films our l'utilisateur *Yasmine* (**droite**) la liste d'amis proposés à *Yasmine*.

3 Démonstration et exemple illustratif

La démonstration de notre système personnalisé de recommandation PERSOREC démontre à travers deux jeux de données du monde réel, *i.e.,* MOVIELENS (*http ://movielens.umn.edu/*)

et BOOKCROSSING(*http ://www.bookcrossing.com/*), le processus de recommandation pour un utilisateur individuel. Il est important de noter que notre système personnalisé de recommandation est générique, *i.e.*, peut être appliqué à n'importe quel jeu de données dont les données ont une structure quadratique (utilisateur, tag, ressource, profil). Dans ce qui suit, nous donnons un exemple illustratif du processus de recommandation sur le jeu de données MOVIELENS. Le processus de recommandation est basé à la fois sur le profil des utilisateurs ainsi que sur les quadri-concepts, *i.e.,* les tags et ressources partagés par des utilisateurs ayant des profils et intérêts équivalents. Considérons l'utilisateur *Yasmine* qui utilise notre système.

La Figure 1 illustre un snapshot (ou instantané) du siteweb PERSOREC pour le jeu de données MOVIELENS. À gauche de la Figure, PERSOREC affiche le profil de *Yasmine* ainsi que les films qu'il a partagé et sa liste d'amis. Tandis qu'au centre de la Figure, nous affichons les recommandations de films avec leurs scores correspondants pour *Yasmine*. À droite de l'interface, nous affichons une proposition d'une liste d'amis pour *Yasmine*. Ainsi, chaque utilisateur du siteweb PERSOREC recevra des recommandations basées sur son profil ainsi que sur les tags et ressources partagés par des utilisateurs ayant le même profil. Par ailleurs, il/elle recevra une liste d'amis proposés. Chaque utilisateur est aussi capable de partager des ressources (recommandés ou pas), d'ajouter des amis ou encore d'annoter des ressources avec des tags (librement choisis ou parmi ceux suggérés). Enfin, avant d'éventuellement ajouter un utilisateur à sa liste d'amis, *Yasmine* a la possibilité de consulter son profil, *i.e.,* ses informations personnelles ou encore ses ressources partagées.

4 Conclusion et Perspectives

Dans ce papier, nous présentons notre système personnalisé de recommandation ainsi qu'une démonstration de notre site web correspondant. En guise de perspectives, nous souhaitons améliorer notre plateforme de recommandations en intégrant des mises à jour incrémentales afin de mettre à jour les recommandations offertes aux utilisateurs.

Références

BELLOGÍN A., CANTADOR I. & CASTELLS P. (2013). A comparative study of heterogeneous item recommendations in social systems. *Inf. Sci.*, **221**, 142–169.

DAS M., THIRUMURUGANATHAN S., AMER-YAHIA S., DAS G. & YU C. (2012). Who tags what ? an analysis framework. *In Proceedings of PVLDB*, **5**(11), 1567–1578.

JELASSI M. N., BEN YAHIA S. & MEPHU NGUIFO E. (2013). A personalized recommender system based on users' information in folksonomies. In *Proc. of the 22nd International Conference on World Wide Web companion*, WWW '13 Companion, p. 1215–1224.

KIM H. K., OH H. Y., GU J. C. & KIM J. K. (2011). Commenders : A recommendation procedure for online book communities. *Electron. Commer. Rec. Appl.*, **10**(5), 501–509.

MIKA P. (2007). Ontologies are us : A unified model of social networks and semantics. *Journal of Web Semantics.*, **5**(1), 5–15.

QUMSIYEH R. & NG Y.-K. (2012). Predicting the ratings of multimedia items for making personalized recommendations. In *SIGIR'12*, p. 475–484, New York, NY, USA : ACM.

F. RICCI, L. ROKACH, B. SHAPIRA & P. B. KANTOR, Eds. (2011). *Recommender Systems Handbook*. Springer.

Système de collecte de données Web pour analyser l'émergence et la propagation de maladies animales

Sylvain Falala[1], Jocelyn De Goër[2], Elena Arsevska[1], Mathieu Roche[3,4], Julien Rabatel[4], David Chavernac[1], Pascal Hendrikx[5], Thierry Lefrancois[1], Barbara Dufour[6], Renaud Lancelot[1]

[1] CIRAD & INRA, UMR CMAEE, Montpellier

[2] INRA, UR EPIA, CLERMONT-FERRAND

[3] CIRAD, UMR TETIS, Montpellier

[4] LABEX NUMEV, LIRMM, Montpellier

[5] ANSES, UCAS, Maisons-Alfort

[6] ENVA, EpiMAI, Maisons-Alfort

Résumé : La veille en santé animale, et notamment la détection précoce d'émergences au niveau mondial d'agents pathogènes, est l'un des moyens permettant de prévenir l'introduction en France de dangers sanitaires (Paquet *et al.*, 2006). Cet article présente une plateforme dédiée à la collecte de données (dépêches) utiles pour la veille automatique. Le recueil des dépêches s'appuie sur des requêtes constituées de mots-clés de maladies, d'hôtes et de symptômes appliquées à Google News. Une interface Web a été développée pour consulter les articles collectés et paramétrer le processus de recueil en définissant de nouvelles combinaisons de mots-clés.
Mots-clés : Veille sanitaire, Collecte de données Web, Recherche d'information

1 Introduction

Dans le cadre de la thématique "Veille sanitaire internationale" de la Plateforme nationale d'épidémiosurveillance en santé animale (Plateforme ESA), le Cirad, l'ANSES et la Direction générale de l'alimentation (DGAl) développent depuis 2013 un système de veille automatique du Web qui effectue :

— le recueil quotidien de dépêches épidémiologiques provenant de sources non officielles, incluant les médias électroniques.
— l'extraction automatique d'informations issues de ces dépêches.
— une restitution synthétique et agrégée de l'information : cartes, séries spatiotemporelles.

Les maladies actuellement surveillées sont la peste porcine africaine, l'Influenza aviaire, la fièvre catarrhale ovine, la fièvre aphteuse et la maladie de Schmallenberg. L'outil est développé de façon générique et permet la surveillance d'autres maladies. Ce système sera utilisé par la Plateforme ESA pour la France et par le réseau de vétérinaires CaribVet situé dans les Caraibes.

2 Méthodes et outils d'acquisition de données Web

Le but de notre système de veille est de disposer d'un outil très réactif qui se veut complémentaire aux sources officielles telles que l'Organisation mondiale de la santé animale (OIE)

ou l'Organisation des Nations unies pour l'alimentation et l'agriculture (FAO).

Le recueil des dépêches s'appuie sur des requêtes constituées de mots-clés de maladies, d'hôtes et de symptômes appliquées à Google News (par exemple, la requête en anglais "*high fever AND mortality AND pigs*" qui combine deux symptômes et un hôte).

Ces mots-clés ont été préalablement définis par des experts et/ou par des méthodes de fouille de textes sur la base du logiciel BioTex (Lossio-Ventura *et al.*, 2016) développé dans le cadre du projet ANR SIFR [1]. BioTex prend en compte deux facteurs pour extraire, de manière automatique, la terminologie dans des corpus textuels. Dans un premier temps, l'approche extrait des termes selon des patrons syntaxiques définis (nom-adjectif, adjectif-nom, nom-préposition-nom, etc.). Après un tel filtrage linguistique, un autre filtrage statistique est appliqué. Celui-ci mesure l'association entre les mots composant un terme. Enfin, des pondérations selon les sources de données sont proposées (Arsevska *et al.*, 2016).

Chaque article est prétraité et normalisé (suppression de balises HTML et Javascript, reconnaissance de la langue, etc.) avant d'être stocké dans une base de données MySQL.

Une interface Web (cf. Figure 1) a été développée en HTML, CSS, PHP et JavaScript. Elle permet de :

— *consulter les articles collectés*. Une section de recherche avancée permet de sélectionner les dépêches en combinant des critères comme les noms de maladie, les hôtes, les symptômes, les noms de source/du média et les dates de publication. Une section "Statistiques" permet de connaître le nombre d'articles recueillis pour une maladie donnée sur une période donnée et d'en observer la distribution dans le temps.

— *paramétrer le processus de recueil* en définissant de nouvelles combinaisons de mots-clés pour Google News.

3 Conclusion et perspectives

Cet article résume les travaux liés au développement d'une plateforme dédiée à la veille automatique allant du recueil des données textuelles (dépêches) jusqu'à la restitution synthétique des informations extraites dans les textes.

L'extraction d'informations sera prochainement intégrée au système. L'extraction dans les dépêches collectées identifie les éléments clés (noms de maladies, lieux, dates, nombres et espèces d'animaux touchés). Elle repose sur des dictionnaires dédiés et des règles préalablement construites par un processus de fouille de données.

Les informations extraites à partir des dépêches seront comparées aux informations issues des données officielles (OIE) afin de mettre en relief la découverte de l'émergence de maladies animales.

1. Semantic Indexing of French Biomedical Data Resources (SIFR) project : `http://www.lirmm.fr/sifr`

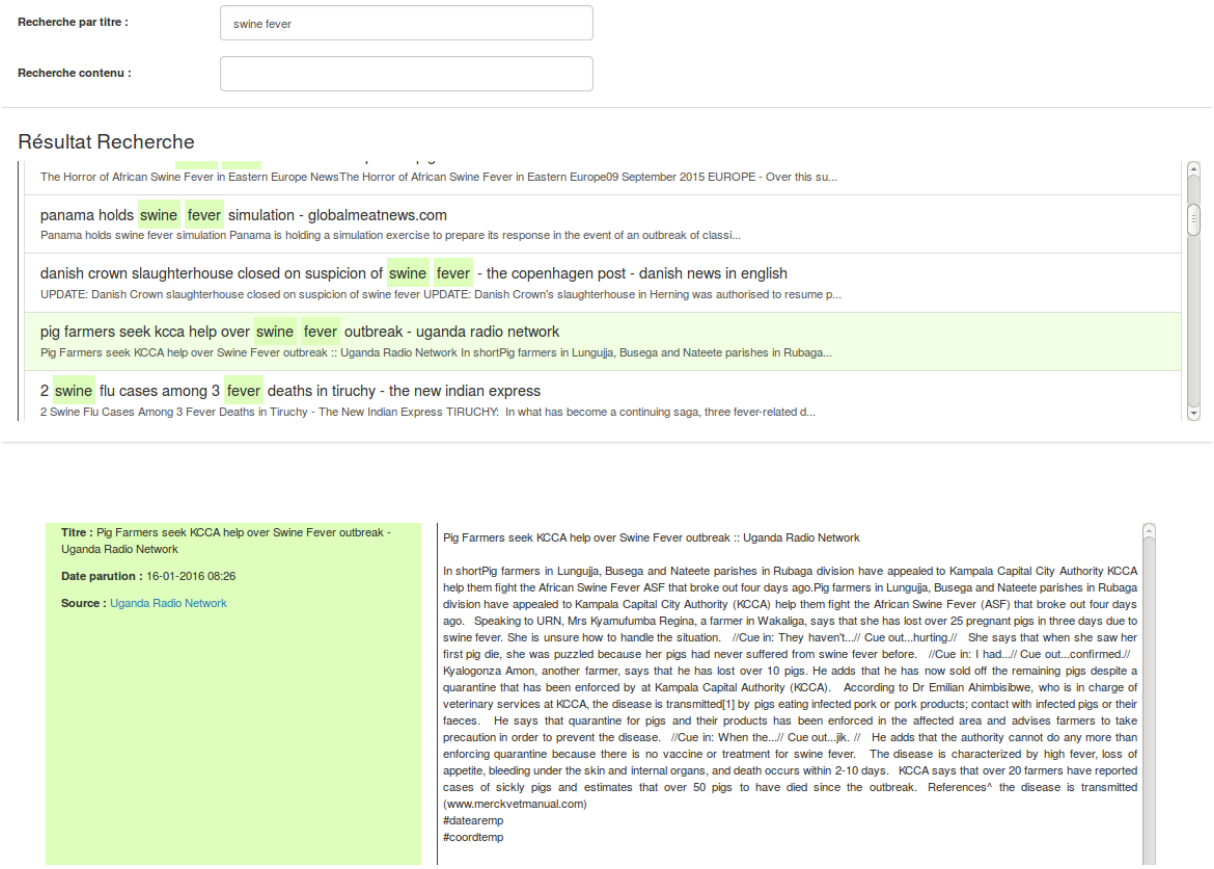


FIGURE 1 – Interface de consultation des dépêches.

Remerciements

Les auteurs remercient les étudiants ayant participé au développement de l'outil : Max Devaud, Thomas Filiol, Baptiste Belot et Clément Hemeury. Ce travail est en partie financé par la DGAl et le Labex Numev [2] (convention ANR-10-LABX-20).

Références

ARSEVSKA E., ROCHE M., HENDRIKX P., CHAVERNAC D., FALALA S., LANCELOT R. & DUFOUR B. (2016). Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web. *Computers and Electronics in Agriculture*, **123**, 104 – 115.

LOSSIO-VENTURA J. A., JONQUET C., ROCHE M. & TEISSEIRE M. (2016). Biomedical term extraction : overview and a new methodology. *Inf. Retr. Journal*, **19**(1-2), 59–99.

PAQUET C., COULOMBIER D., KAISER R. & CIOTTI M. (2006). Epidemic intelligence : a new framework for strengthening disease surveillance in europe. *Euro surveillance*, **11**(12), 212–214.

2. http://www.lirmm.fr/numev

Structuration sémantique des sensations tactiles

Bruno Albert[1,2,3], Cecilia Zanni-Merk[1], François de Bertrand de Beuvron[1], Jean-Luc Maire[2], Maurice Pillet[2], Julien Charrier[3], Christophe Knecht[3]

[1] Laboratoire ICube, Insa de Strasbourg, Illkirch
bruno.albert@ineva.fr
[2] Laboratoire SYMME, Université Savoie Mont-Blanc, Annecy le Vieux
[3] Société INEVA, Illkirch

Résumé : Le toucher est largement impliqué dans le jugement et le choix d'un produit par les consommateurs. En revanche, dans l'industrie, la formalisation du contrôle qualité lié à ce sens est souvent inexistante, ce qui cause généralement de fortes variabilités de contrôle. L'objectif de cette étude est donc de proposer une méthodologie structurée et générique pour le contrôle qualité tactile, ainsi que de permettre l'automatisation de ce type de contrôle. Une description générique des sensations tactiles est proposée. Elle se compose de catégories de sensations élémentaires et a été construite sur des bases sémantiques ce qui lui confère un caractère générique. Une représentation ontologique de haut niveau est également proposée afin de structurer les concepts mises en jeux dans l'évaluation de la qualité tactile, de représenter les relations existantes entre les descripteurs tactiles, et permettra par la suite de faire le lien avec des mesures obtenues grâce à des capteurs.

Mots-clés : Qualité perçue, analyse sémantique, sensations tactiles, ontologie tactile

1 Introduction

La qualité perçue est un facteur important du jugement et du choix d'un produit par les consommateurs. Différents sens sont impliqués, et en particulier le sens du toucher. En revanche, à l'inverse de la vision, peu d'études se sont intéressées au toucher dans le domaine du contrôle de la qualité des produits, sens qui est pourtant largement utilisé. Il y a en effet un fort besoin de structuration et de généralisation de ce type de contrôle au sein de l'industrie afin de réduire la variabilité du contrôle.

Le toucher fait appel à des phénomènes physiques complexes, principalement liés au contact et à la friction. Maîtriser ce type de contrôle suppose donc de comprendre comment s'élabore la perception tactile d'un produit. Le contrôle de la qualité par rapport à un référentiel apparaît important pour garantir la satisfaction du client. Si l'on examine les descripteurs tactiles qui sont classiquement proposés, on s'aperçoit que ceux-ci sont généralement très liés au produit examiné et au langage utilisé. De plus, ils ont très souvent vocation à être utilisés dans le cadre d'une démarche de conception de nouveaux produits.

L'objectif de cette étude est de proposer une méthodologie structurée et générique pour le contrôle qualité tactile, ainsi que de permettre l'automatisation de ce type de contrôle. Une analyse des descripteurs usuels des perceptions tactiles, ainsi que des méthodes de classifications a été réalisée afin de proposer une description générique des sensations tactiles. Celle-ci est composée de catégories de sensations élémentaires. Une représentation ontologique de haut niveau est également proposée afin de répondre à la problématique de structure des connaissances liées au tactile. A terme, elle permettra de représenter la description des sensations tactiles ainsi que les relations entre les descripteurs, et par la suite de faire le lien avec des mesures obtenues grâce à des capteurs.

2 Identification de sensations tactiles élémentaires

Les descripteurs trouvés dans la littérature sont généralement liés à des types de produits et matériaux spécifiques. Ils dépendent également du langage utilisé et de la culture des contrôleurs. En partant de tous les descripteurs de sensations tactiles répertoriés dans la littérature – on peut citer par exemple les études suivantes : (Crochemore, Vergneault, & Nesa, 2003; Dumenil-lefebvre & Int, 2006; Giboreau et al., 2007) – ainsi que dans des bases de données et dictionnaires, nous nous sommes ici intéressés à la sémantique de ces mots. Plus de 200 descripteurs ont ainsi été analysés en utilisant des méthodes de classification et les relations sémantiques de type synonymes et antonymes. Les descripteurs trouvés ont en particulier été classés suivant les axes "source", "effet" et "propriété physique" (David, 1997). Ces axes mettent en évidence les bases sémantiques des descripteurs et les liens avec les caractéristiques d'une surface.

Les relations de synonymes et antonymes entre les descripteurs tactiles usuels peuvent être utilisées pour tracer des correspondances sémantiques entre ces descripteurs. Associée à un outil graphique, la méthode OpenOrd (Martin, Brown, Klavans, & Boyack, 2011) a été utilisée afin de positionner graphiquement les descripteurs suivant leur relations et d'extraire des groupement sémantiques. Ces groupements ont ainsi servi de base à l'élaboration de sensations élémentaires : *Adhérence*, *Relief*, *Dureté*, *Réactivité*, *Trace*, *Chaleur*, *Douleur*, *Poids*, *Uniformité* et *Forme*. Chacun des descripteurs de la littérature peut ainsi relever d'une ou plusieurs sensations élémentaires.

3 Introduction d'une représentation ontologique

La conceptualisation des connaissances est généralement une approche très appropriée dans le but de formaliser un domaine d'intérêt. En particulier, l'utilisation d'ontologies est pertinente en considérant des données et connaissances sémantiques (Zanni-Merk, 2015). Même si leur utilisation est devenue relativement standard dans les systèmes à base de connaissances, peu d'études ont proposé des ontologies directement en lien avec la description des perceptions humaines, ou plus spécifiquement avec la qualité perçue. Dans un objectif d'automatisation du processus de contrôle de qualité tactile, le problème de l'ancrage des symboles doit être abordé (Coradeschi & Saffiotti, 2003). Il s'agit en particulier d'établir les correspondances entre les données de capteurs et les symboles qui sont ici les sensations tactiles. Certains exemples d'ontologies développées dans d'autres domaines montrent leur intérêt et leur potentiel pour ce type d'applications (Roda & Zanni-Merk, 2016; Zanni-Merk, Marc-Zwecker, Wemmert, & Bertrand de Beuvron, 2015). Un exemple rare d'ontologie proposé dans un domaine similaire par Myrgioti (2013) est focalisé sur le développement logiciel d'interfaces haptique.

La première étape de construction d'une représentation conceptuelle d'un domaine est l'emploi d'une ontologie de haut niveau afin de former le "squelette" de la structure. Il existe de multiples ontologies de haut niveau, et nous nous concentrerons ici sur l'ontologie Semantic Sensor Network (SSN) proposée par Compton et al. (2012), qui a été identifiée comme étant particulièrement pertinente en considérant le contexte de l'étude et les perspectives de développements futurs.

Compton et al. (2012) ont introduit l'ontologie SSN dans le but de décrire les capteurs et observations. Outre les possibilités de développements futurs autour du système de capteurs, l'ontologie SSN propose une manière de conceptualiser les liens entre les propriétés, les capteurs et le contexte (observations). En effet, cette représentation rassemble les concepts nécessaires à la formation de la perception tactile et les structure de manière à permettre l'intégration de connaissances provenant de ce domaine. Les concepts de cette ontologie étant de suffisamment haut niveau, ils peuvent être alignés au domaine des sensations tactiles. En considérant uniquement la description de ces sensations, les concepts de *Property* et *FeatureOfInterest* peuvent être respectivement alignés aux concepts de sensations élémentaires et aux descripteurs de la littérature. De plus, les relations entre ces entités peuvent être dérivées

des connaissances d'experts, et utilisées afin de donner un sens aux mesures provenant de capteurs à propos des perceptions tactiles.

4 Conclusion

Les recherches menées durant la première phase de ce projet ont permis la compréhension du processus de perception tactile humain et le développement d'une description générique et structurée des sensations tactiles. Ce faisant, les descripteurs usuels de la perception tactile ont été listés et leur spécificités expliquées. Ils ont été classés suivant différents modes afin d'extraire des groupements sémantiques et de construire une grille de description des sensations tactile à partir de catégories élémentaires, qui se veut minimale et complète. Finalement, une première représentation conceptuelle du problème de description tactile a été introduite, basée sur le principe d'ontologie. Un plan de validation permettra par la suite de vérifier l'exhaustivité et la pertinence de la méthode, ainsi que d'évaluer ses performances et amener à de possibles améliorations.

Ce travail pose les bases de la formalisation du contrôle qualité tactile et fournira les éléments nécessaires à l'évaluation des sensations tactiles dans un contexte de production industrielle. Cette méthode permettra à terme de réduire les variabilités du contrôle liées à la subjectivité des opérateurs et au manque de structure actuel, et rendra possible le développement de systèmes automatisés pour ce type de contrôles.

Références

Compton, M., Barnaghi, P., Bermudez, L., García-Castro, R., Corcho, O., Cox, S., … Taylor, K. (2012). The SSN ontology of the W3C semantic sensor network incubator group. *Journal of Web Semantics*, *17*, 25–32. http://doi.org/10.1016/j.websem.2012.05.003

Coradeschi, S., & Saffiotti, A. (2003). An introduction to the anchoring problem. *Robotics and Autonomous Systems*, *43*(2-3), 83. http://doi.org/10.1016/S0921-8890(03)00020-4

Crochemore, S., Vergneault, C., & Nesa, D. (2003). A new reference frame for tactile perceptions: Sensotact. *5th Rose Mary Pangborn, Boston MA, USA*, 20–24.

David, S. (1997). Représentations sensorielles et marques de la personne : contrastes entre olfaction et audition. *Catégorisation et Cognition: De La Perception Au Discours. D. Dubois. Paris, Kimé*, 211–242.

Dumenil-lefebvre, A., & Int, A. D. (2006). Int ´ egration des aspects sensoriels dans la conception des emballages en verre : mise au point d ' un instrument m ´ ethodologique ` a partir des techniques d ' ´ evaluation sensorielle.

Giboreau, A., Dacremont, C., Egoroff, C., Guerrand, S., Urdapilleta, I., Candel, D., & Dubois, D. (2007). Defining sensory descriptors: Towards writing guidelines based on terminology. *Food Quality and Preference*, *18*(2), 265–274. http://doi.org/10.1016/j.foodqual.2005.12.003

Martin, S., Brown, W. M., Klavans, R., & Boyack, K. W. (2011). OpenOrd: an open-source toolbox for large graph layout. *Proc. SPIE*, *7868*(JANUARY 2011), 786806–786811. http://doi.org/10.1117/12.871402

Myrgioti, E., Bassiliades, N., & Miliou, A. (2013). Bridging the HASM: An OWL ontology for modeling the information pathways in haptic interfaces software. *Expert Systems with Applications*, *40*(4), 1358–1371. http://doi.org/10.1016/j.eswa.2012.08.053

Roda, F., & Zanni-Merk, C. (2016). An Intelligent Data Analysis framework for supporting perception of geospatial phenomena. In I. (International A. for Ontology & its Applications) (Eds.), *FOIS 2016 (9th International Conference on Formal Ontology in Information Systems)*.

Zanni-Merk, C. (2015). KREM : A Generic Knowledge-Based Framework for Problem Solving in Engineering Proposal and Case Studies. In *Keod* (pp. 381–388).

Zanni-Merk, C., Marc-Zwecker, S., Wemmert, C., & Bertrand de Beuvron, F. de. (2015). A Layered Architecture for a Fuzzy Semantic Approach for Satellite Image Analysis. *International Journal of Knowledge and Systems Science*, *6*(2), 31–56. http://doi.org/10.4018/IJKSS.2015040103

Représentation holistique des projets de construction à l'aide de modèles relationnels probabilistes

C. Baudrit, T.T.P. Tran, F. Taillandier, D. Breysse

INRA, I2M, USC 1368, F-33400 Talence, France.
cedric.baudrit@u-bordeaux.fr,thi-thuy-phuong.tran@u-bordeaux.fr,
franck.taillandier@u-bordeaux.fr,denis.bresse@u-bordeaux.fr

Résumé : Les facteurs humains associés aux risques naturels peuvent être sources de défaillances à plus ou moins long terme dans les projets de construction à différents niveaux d'échelles. Bien qu'il existe une littérature substantielle sur le management des risques dans les projets de construction, aucune étude ne propose une approche générique et holistique capable de représenter le système dans sa globalité tout en prenant en compte les incertitudes. Ce papier propose un modèle de connaissance générique basé sur les modèles relationnels probabilistes (couplant ontologie et réseaux Bayésiens). Pour illustrer notre approche, le modèle est instancié dans le cadre d'un projet de rénovation d'une route et d'un pont situé au Vietnam.

Mots-clés : Ontologie, Réseaux Bayésiens, Modèle relationnel probabiliste, incertitude.

1 Introduction

La construction d'un ouvrage doit répondre à des objectifs économiques (coût de l'ouvrage), temporels (délai de réalisation) et de performance (qualité de l'ouvrage livré). Mais les erreurs humaines et les aléas qu'ils soient naturels ou technologiques, peuvent être sources de défaillances empêchant d'atteindre les objectifs fixés. Face à cette complexité, il est capital de proposer des approches conceptuelles qui cherchent à représenter l'ensemble des acteurs et des éléments fonctionnels en interactions à différents niveaux d'échelles dont le comportement collectif engendre des structures organisées qui influencent en retour des comportements individuels. En effet, comprendre indépendamment le comportement de chacun des constituants du système n'est pas suffisant pour comprendre le système dans son ensemble. Cette mutualisation des savoirs a pour vocation de mieux appréhender le système dans sa globalité et ainsi de gérer efficacement les risques. Bien qu'il existe une littérature substantiel sur le management des risques dans les projets de construction Diraby *et al.* (2005); Zhong & Li (2014); Mehdizadeh (2012); Zhang *et al.* (2014), aucune étude ne propose d'approche holistique capable de représenter et simuler un projet de construction avec prise en compte des incertitudes. Nous devons être capable de structurer, de représenter et de mutualiser des connaissances hétérogènes entachées d'incertitudes (Dubois, 2007) issues de différentes sources et de différentes disciplines dans un cadre formel unificateur. Dans ce contexte, le cadre des modèles relationnels probabilistes (PRMs) (Getoor *et al.*, 2007) fournit un formalisme mathématique qui permet de décrire des systèmes dynamiques stochastiques. Le concept des PRMs est basé sur une extension des réseaux Bayésiens (Pearl, 1988) qui est couplé à un réseau sémantique du domaine de connaissance (Guizzardi, 2005). Le papier propose une structure holistique des projets de construction à l'aide des PRMs et instancie un modèle dans le cadre d'un projet réel de rénovation de route et de ponts au Vietnam.

2 Modèles relationnels probabilistes PRMs

Une ontologie légère (Guizzardi, 2005) peut être définie par un 4-tuple $< \mathcal{C}, \mathcal{R}, \mathcal{A}, \mathcal{I} >$ où $\mathcal{C}$, $\mathcal{R}$, $\mathcal{A}$, $\mathcal{I}$ sont des ensembles disjoints contenant les concepts, les relations, les attributs et les instances. Un PRM est un graph acyclique dirigé dont les noeuds $C.A$ (en adoptant le paradigme objet) représentent des variables aléatoires et dont les arcs encodent des dépendances entre les variables. Le formalisme des PRMs induit un réseau Bayésien et définit une probabilité jointe

$$P(\mathcal{I}) = \prod_{C \in \mathcal{C}} \prod_{A \in \mathcal{A}(C)} \prod_{c \in \sigma_{\mathcal{R}}(C)} P(c.A|\text{Pa}(c.A)) \quad (1)$$

sur le monde des possibles instancié par un squelette, dénoté $\sigma_{\mathcal{R}}$, spécifiant l'ensemble des objets $\sigma_{\mathcal{R}}(C)$ pour chaque concept et les relations entre les concepts, *i.e* $\sigma_{\mathcal{R}}(C)$=$< \mathcal{I}(C), \mathcal{R}(C) >$ où Pa$(c.A)$ correspond à l'ensemble des parents de $c.A$ et $P(c.A|\text{Pa}(c.A))$ la probabilté conditionelle de $c.A|\text{Pa}(c.A)$. La différence entre $\sigma_{\mathcal{R}}$ et l'instantiation d'un modèle ontologique vient du fait qu'aucune valeur n'est affectée aux attributs $c.A$.

3 Représention holistique des projets de construction

La figure 1 montre la structure du modèle PRMs construit à partir de la littérature et d'expertise fournissant une vision holistique du système des projets de construction. Basée sur une taxo-

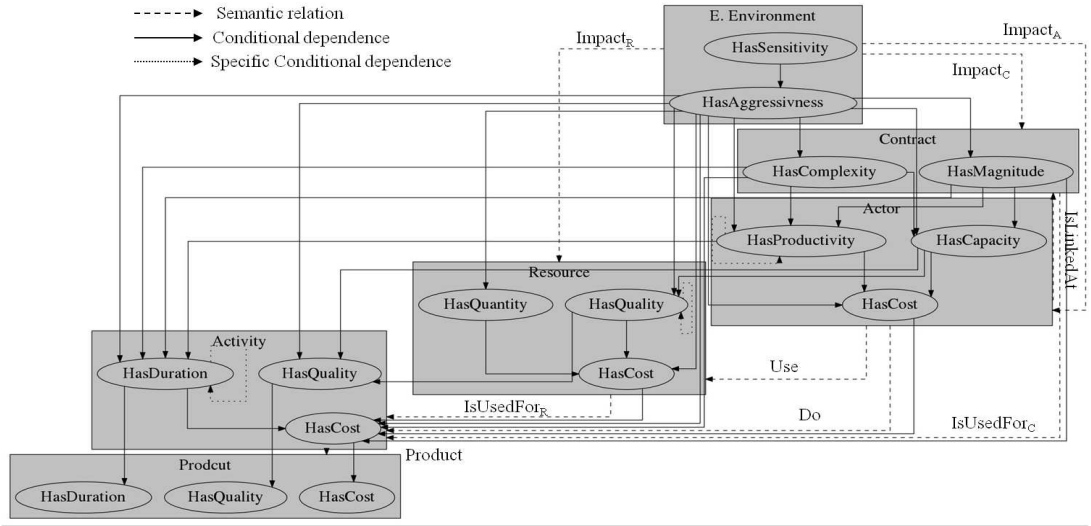


FIGURE 1 – Modèle générique basé sur le formalisme des PRMs pour représenter les projets de construction

nomie développée par Niu & Issa (2015), 6 concepts ont été retenues pour décrire le domaine à un haut niveau de description (voir Fig. 1). Chaque concept comme $Actor$ (*resp.* $Resource$) est caractérisé par des variables aléatoires (*i.e* attributs incertains) tels que $HasCapacity$ qui exprime sa capacité pour exécuter une tache (*resp.* $HasCost$ résumant le coût d'une commande et du stockage). A titre d'exemple, les acteurs sont reliés par des contrats pour faire

une activité menant ainsi à définir les relations d'associations $IsLinkedAt$, Use, Do (voir Fig. 1). Le contenu des contrats peut changer selon les conditions environnementales EE qui peut se caractériser par une dépendance entre les variables $EE.HasAggressiveness$ et $Contract.HasComplexity$ (voir Fig 1). La structure du modèle a été implémentée avec le langage O3PRM dans aGrUM (agrum.lip6.fr) qui utilise une syntaxe orienté objet (Torti *et al.*, 2010). Une instatiation du modèle a été générée pour décrire un projet de rénovation d'une route et d'un pont au Vietnam. Durant le déroulement réel du projet, différents évènements (une tempête suivie d'une inondation, des accidents de personnes, des machines cassées et une augmentation de salaires) ont induit des retards notables (2 mois) ainsi qu'un surcoût. La simulation du modèle instancié, en considérant le scénario réel (aléas météorologiques, accidents...) a prédit un retard du projet avec une confiance de 60% ; ce qui est concordant avec la réalité.

4 conclusion

Un modèle holistique générique permettant de formaliser la connaissance et de simuler le risque dans les projets de construction a été proposé à l'aide du formalisme des PRMs. Un cas réel simplifié a été instancié et les premiers résultats de simulations font sens avec le retard du projet. La prochaine étape consistera à rendre le modèle plus spécifique de par l'utilisation des relations sémantiques de type taxonomique et méréologique, et ainsi affiner les échelles de description du modèle générique. La généricité du modèle devra être testée sur d'autres projets de construction.

Références

DIRABY E. T. A., LIMA C. & FEIS B. (2005). Domain Taxonomy for Construction Concepts : Toward a Formal Ontology for Construction Knowledge. *J COMPUT CIVIL ENG*, **19**(4), 394–406.

DUBOIS D. (2007). Uncertainty theories : a unified view. In *Cybernetic Systems, Dublin (Ireland)*, p. 4–94, http ://www.ieee.org/ : IEEE.

GETOOR L., FRIEDMAN N., KOLLER D., PFEFFER A. & TASKAR B. (2007). Probabilistic relational models. In L. GETOOR & B. TASKAR, Eds., *Introduction to Statistical Relational Learning*.

GUIZZARDI G. (2005). *Ontological foundations for structural conceptual models*. PhD thesis, Enschede.

MEHDIZADEH R. (2012). *Dynamic and multi-perspective risk management of construction projects using tailor-made Risk Breakdown Structures*. PhD thesis.

NIU J. & ISSA R. R. (2015). Developing taxonomy for the domain ontology of construction contractual semantics : A case study on the {AIA} {A201} document. *ADV. ENG. INFORM.*, **29**(3), 472 – 482.

PEARL J. (1988). *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

TORTI L., WUILLEMIN P.-H. & GONZALES C. (2010). Reinforcing the Object-Oriented Aspect of Probabilistic Relational Models. In *Proceedings of the 5th PGM*, p. 273–280.

ZHANG L., WU X., SKIBNIEWSKI M.J. Z. J. & Y. L. (2014). Bayesian-network-based safety risk analysis in construction project. *RELIAB ENG SYST SAFE*, **131**, 29–39.

ZHONG B. & LI Y. (2014). An ontological and semantic approach for the construction risk inferring and application. *Journal of Intelligent & Robotic Systems*, **79**(3), 449–463.

Construction d'un *gold standard* pour des données agronomiques

Imène Chentli[1,2,3], Pierre Larmande[3] et Konstantin Todorov[1,2]

[1] Université de Montpellier,
chentli.imene@hotmail.fr
[2] Laboratoire de Robotique, d'Informatique et Micro-électronique de Montpellier (LIRMM),
Konstantin.Todorov@lirmm.fr
[3] Institut de Biologie Computationnelle (IBC), Montpellier
Pierre.Larmande@ird.fr

Résumé : Dans le contexte des ressources agronomiques et au jour d'aujourd'hui, il n'existe ni de données de référence ni de questions adaptées au domaine pour évaluer les interfaces dotées de Systèmes à Questions-Réponses (SQR). A cet effet, nous avons construit un étalon-or (*gold standard*) constitué d'un ensemble de questions de référence formulées en langage naturel et de leur correspondance en langage SPARQL. Nous considérons la base de connaissances d'Agronomic LinkedData comme le jeu de données de référence.

Mots-clés : Données de référence, Web sémantique, Agronomie, Interrogation de données, Systèmes de Questions-Réponses, Traitement du Langage Naturel, SPARQL.

La nouvelle génération de web de données liées a pour but de lever les difficultés liées à la recherche d'information grâce à une représentation sémantique des connaissances stockées en format RDF[1]. En biologie par exemple, les principales problématiques sont de comprendre comment les experts vont pouvoir accéder à ces données et comment être sûr qu'il n'en manque pas dans les résultats proposés vu la diversité et la multiplicité des ressources. En effet, les sciences de la vie dans le web de données liées comprennent 192 millions de liens pointant vers d'autres domaines et plus de 3 milliards de triplets RDF soit environ 10% du nombre total de triplets de tous les domaines confondus. En trois années, l'ensemble de données passe de 41 à 85 ensembles [7]. Ces ressources structurées sont accessibles et mises à jour régulièrement pour leur majorité *via* différentes Interfaces à Langage Naturel (ILN). Ces dernières permettent aux experts de profiter de l'expressivité puissante des standards du Web Sémantique (WS) sans se soucier de leur complexité. Ainsi, les biologistes expriment leurs besoins d'information en Langage Naturel (LN). Puis, l'ILN les *traduit* en un langage formel. Et lorsqu'il est possible, la machine propose une réponse adaptée. A cet effet, le passage du LN au langage de machine nécessite une étape intermédiaire de *traduction* ou de *formalisation*. L'accès aux connaissances reste une problématique non résolue dépendant de l'expertise du biologiste. Le langage SPARQL[2], standard recommandé par le W3C[3], est utilisé comme langage formel pour l'interrogation, l'accès aux bases de connaissances et *in fine*, la traduction du LN. Cependant, sa complexité peut restreindre son utilisation. Ainsi, la semi-automatisation de la traduction des requêtes en LN vers le SPARQL fait l'objet de recherche dans différents domaines. L'intérêt est d'éviter aux biologistes et autres utilisateurs non-spécialistes, la confrontation à la complexité du SPARQL notamment pour sa syntaxe et l'organisation des données dans les bases de connaissances.

L'objectif de notre travail est double : établir une analyse empirique d'ILN pouvant potentiellement être exploitées dans le domaine agronomique puis, construire un *gold standard* dédié à la communauté agronomique car jusqu'à présent, il n'existe ni de données de références

[1] RDF : Ressource Description Framework est un modèle de graphe décrivant les ressources du Web sous forme de triplets (Sujet, Prédicat, Objet) et un langage de base du Web Sémantique
[2] SPARQL : SPARQL Protocol and RDF Query Language
[3] W3C : World Wide Web Consortium. Accès *via* ce lien : http://www.w3.org

ni de questions préétablies pour l'évaluation des systèmes les exploitant. Nous proposons ainsi les ressources d'Agronomic LinkedData (AgroLD) comme jeu de données de référence et un ensemble de 50 questions construites en LN et leur correspondance en langage SPARQL. Cette première étape permettra la préparation aux phases de tests et d'évaluation des ILN.

1 Travaux existants

Les ILN qui emploient des Systèmes à Questions-Réponses (SQR), dépendent du degré d'expertise pour le domaine et le langage formel tel que le SPARQL. D'autres ILN dépendent de la liberté de leur utilisation pouvant aller d'une interface proposant un formulaire ou un simple champ textuel jusqu'aux interfaces les plus complexes où l'information est visualisée dynamiquement. L'architecture d'un SQR est généralement composée de trois modules. Le premier analyse la question, le deuxième sélectionne un ensemble de résultats candidats et le dernier analyse les candidats et en extrait les réponses les plus pertinentes si elles sont présentes.

Notre attention est portée sur des SQR utilisant une voire plusieurs approches de traduction du LN en requêtes SPARQL pour interroger des bases de connaissances structurées. Notre champ d'étude considère tous les types de requêtes (mots clés, fragments ou phrases intégrales) et il est élargi à tous les domaines d'application car au jour d'aujourd'hui, il existe très peu de SQR dédiés au domaine biologique [6]. Des travaux de traduction du LN vers des requêtes sémantiques ont mis en évidence l'emploi de différentes approches. Nous en énumèrerons les principales en nous limitant à un exemple d'ILN par approche :

a) **Approche basée sur des requêtes en LN contrôlé** - GINSENG[1] peut contrôler les requêtes selon l'ontologie fournie à l'outil. Ginseng utilise des règles statiques définissant la structure générale de la requête et dynamiques issues des étiquettes de la base de connaissances. Cependant, l'expert est limité dans la manière de rédiger ses requêtes. Lorsque le terme n'apparaît pas dans la liste déroulante cela signifie que le système considère que sa grammaire est incorrecte et ne l'accepte pas.
b) **Approche basée sur l'alignement de structures sémantiques aux ontologies** – QUERIX[4] indépendant du domaine, analyse la structure syntaxique de la requête pour trouver de meilleures correspondances avec la base de connaissances. Une fenêtre de dialogue propose différents triplets à l'utilisateur. Lorsque l'utilisateur les sélectionne, une requête en SPARQL est générée et exécutée. Cependant, le système ne résout pas les ambiguïtés. Les requêtes sont en Anglais et doivent impérativement commencer par « Wh ».
c) **Approche basée sur une grammaire formelle** - ORAKEL[2] utilise deux types de lexiques lors du processus d'analyse. Le premier est général et indépendant du domaine pour la construction de la requête. Le second correspond au domaine. ORAKEL a donc besoin de connaissances ontologiques pour interpréter sémantiquement des termes ou pour résoudre des ambiguïtés. Si la source ontologique est incomplète, les résultats le seront aussi lors de son évaluation.
d) **Approche basée sur des patrons** - LODQA[5] analyse les requêtes d'un point de vue linguistique, aligne les phrases nominales aux différents termes ontologiques et propose des résultats en SPARQL. L'avantage repose sur sa modulabilité pour analyser la requête. Par ailleurs, le prototype reste en cours de développement et limité à une ontologie et au domaine.
e) **Approche basée sur l'auto-génération de requêtes SPARQL** - SPARKLIS[3] est indépendant du domaine. L'outil permet d'exploiter différents types de bases de connaissances *via* un point d'accès SPARQL en guidant l'utilisateur pas à pas à construire les questions, les requêtes en SPARQL et obtenir les réponses de façon interactive. Le système donne également des suggestions pour raffiner la sélection.

Plusieurs campagnes d'évaluation de SQR sont menées régulièrement telles que QALD[4] et BioNLP[5] pour le domaine biomédical. Par ailleurs, les challenges de SQR sont inexistants pour les ressources agronomiques. Chaque domaine étant caractérisé par ses propres terminologies il serait alors difficile de tester un SQR avec un même jeu de requêtes. Le portage de SQR vers le domaine agronomique nécessite donc la mise en place et de l'analyse d'un nouveau jeu de requêtes.

2 Démarche

Notre travail consiste à mettre en place un *gold standard* à la disposition de la communauté des plantes et de faciliter l'évaluation des SQR en respectant les besoins exprimés par les experts et leur capacité à faire face aux grands volumes hétérogènes de données. L'objectif final consiste à ce que le modèle à tester retourne, pour une question en LN et une source de données RDF, une liste d'entrées répondant aux questions posées. Pour ce faire, nous avons tenu compte des éléments suivants :

a) **L'ensemble de données**. Ce sont des données interconnectées et structurées en graphes RDF. Nous exploiterons la base de connaissances du projet AgroLD[8]. En effet, elle est accessible à l'ensemble de la communauté et représente une grande diversité de données agronomiques.
b) **L'infrastructure**. Dans notre cas, elle correspond au point d'accès SPARQL d'AgroLD consultable directement en ligne.
c) **Un *gold standard*** permettant d'effectuer des tests d'entraînement à l'aide de l'ensemble de données. Pour ce faire, nous avons construit manuellement 50 questions en LN et 50 requêtes en SPARQL correspondantes. Ces dernières sont soumises au point d'accès SPARQL d'AgroLD pour extraire les réponses correspondantes. Une fois validé, ce *gold standard* servira *in fine* de référence. Des exemples de requêtes sont présentés *via* le lien suivant : http://volvestre.cirad.fr:8080/aldp/sparqleditor.jsp
d) **Un modèle** correspond à un voire une liste de SQR à évaluer selon une procédure et des mesures d'évaluation.

Référence

[1] BERNSTEIN A., KAUFMANN E., KISER C., KIEFER C. (2006). Ginseng: A Guided Input Natural Language Search Engine for Querying Ontologies, 2006 Jena User Conference, Bristol, UK, May 2006.

[2] CIMIANO P., HAASE P. & HEIZMANN J. (2007). Porting natural language interfaces between domains: An experimental user study with the orakel system. In Proceedings of the 12th international conference on intelligent user interfaces (pp. 180-189). ACM.

[3] FERRÉ S. (2014). SPARKLIS: a SPARQL endpoint explorer for expressive question answering. In ISWC posters & demonstrations track (Vol. 1272, pp. 45-48).

[4] KAUFMANN E., BERNSTEIN A. & ZUMSTEIN R. (2006). Querix: A natural language interface to query ontologies based on clarification dialogs. Springer.

[5] KIM J.D. & COHEN K. (2013). Natural language query processing for sparql generation: A prototype system for SNOMEDCT. In Proceedings of the BioLINK SIG (pp. 32-38).

[6] NEVES M. & LESER U. (2015). Question answering for Biology. Methods, 74, 36-46.

[7] SCHMACHTENBERG M., BIZER C., PAULHEIM H. (2014) Adoption of the Linked Data Best Practices in Different Topical Domains. 13th International Semantic Web Conference (ISWC2014) – RDB Track (2014)

[8] VENKATESAN A., LARMANDE P., JONQUET C., RUIZ M. & VALDURIEZ P. (2015). Facilitating efficient knowledge management and discovery in the Agronomic Sciences. In Proceedings of the 8th Semantic Web Applications and Tools

[4] QALD : http://www.sc.cit-ec.uni-bielefeld.de/qald
[5] BioNLP : http://2016.bionlp-st.org

Interopérabilité sémantique dans le domaine du diagnostic in vitro : évaluation d'algorithmes sur LOINC® et l'ontologie SNOMED CT®

Mélissa Mary[1, 2], Lina F. Soualmia[2, 3] et Xavier Gansel[1]

[1] bioMérieux SA, Département Développement et Intégration,
38390 La Balme Les Grottes
{melissa.mary, xavier.gansel}@biomerieux.com
http://biomerieux.com
[2] LITIS EA 4108 et NormaSTIC CNRS 3638, Université de Normandie,
Lina.Soualmia@chu-rouen.fr
[3] LIMICS INSERM UMR_1142, Sorbonne Universités, Paris.

Résumé : Nous proposons dans cet article des méthodes d'évaluation d'algorithmes d'alignement entre Systèmes d'Organisation des Connaissances (SOC) de référence représentant le domaine du diagnostic in vitro. Les méthodes proposées reposent sur trois mesures de similarité syntaxique et un algorithme à base d'heuristiques. Les résultats que nous obtenons dans cette étude montrent que les métriques de similarité syntaxique ne se révèlent pas suffisamment probantes pour se voir appliquées de manière systématique au domaine des tests de laboratoire. En revanche, la qualité des alignements obtenus via l'algorithme heuristique, filtré a posteriori en fonction d'une dimension sémantique, permettent de conforter les critères de performance que nous avions établis.

Mots-clés : algorithme d'alignement, domaine de la santé, évaluation, interopérabilité sémantique, systèmes d'organisation des connaissances.

1 Introduction

La centralisation des données du patient dans un répertoire électronique est gérée par les instituts de santé publique afin d'améliorer la prise en charge du patient et de maîtriser les coûts médicaux (Macary, 2007). L'interopérabilité entre les différents systèmes d'information utilisés par les professionnels de santé est un enjeu majeur dans la mise en place de ces dossiers électroniques qui sont codés à l'aide de Système d'Organisation des Connaissances (SOC) (Hodge, 2000). Dans le domaine du diagnostic *in vitro* deux SOC sont préconisés pour coder les informations relatives à la description des tests (LOINC®) et à l'expression des résultats obtenus (SNOMED CT®). L'obtention d'alignements entre SOC est un enjeu majeur tant du fait de la volumétrie des données à aligner que de la qualité attendue des alignements obtenus par les algorithmes. De nombreuses méthodes ont été développées permettant de réaliser des alignements notamment entre ontologies (Euzenat & Shvaiko, 2013). Dans cet article, notre étude vise à évaluer des métriques d'alignement, syntaxiques et sémantique sur des termes extraits de LOINC® et SNOMED CT®. L'évaluation est réalisée grâce à un jeu d'alignements entre LOINC® et SNOMED CT® (IHTSDO & Regenstrief Institute, 2013) sont le résultat d'une collaboration initiée en 2013 entre l'IHTSDO et le *Regenstrief Institute*.

2 Matériel et Méthode

Alignement LOINC® SNOMED CT®

Dans cette étude nous utilisons la première version d'alignement publiée en septembre 2014 (2 177 alignements), qui couvre 0,15% des tests LOINC® et 2,115 parties (IHTSDO et Regenstrief Institute, 2013).

Similarités syntaxiques

La similarité syntaxique est un score calculé par comparaison des chaînes de caractères. Nous avons étudié trois méthodes de calcul de similarité:

- Distance de Damereau Levenshtein (Damerau, 1964; Levenshtein, 1966) ;
- Similarité de Stoilos (Stoilos *et al.*, 2005) ;
- Similarité WGram qui permet de s'affranchir de l'ordre des mots composant un terme.

Similarité sémantique

La similarité sémantique est obtenue à partir du Metathesaurus® (Fung & Bodenreider, 2005) et construit à partir de 195 ressources biomédicales ; les 13 millions de termes sont représentés par 3 millions de concepts identifiés de manière unique par un code (CUI). Nous avons développé un algorithme basé sur l'outil MetaMap qui permet d'identifier les CUIs correspondants à chacun des termes d'un SOC. La similarité sémantique est calculée de manière indirecte par comparaison des CUIs représentant les termes à aligner et un indice de confiance de l'alignement entre le terme et les CUIs du Metathesaurus®.

Alignement heuristique par la méthode des ancres

La méthode des ancres que nous proposons est une stratégie d'alignement heuristique, initiée à partir d'un alignement préexistant entre les deux SOC qui s'inspire du travail de Seddiqui et Aono (Seddiqui & Aono, 2009). Le principe de la méthode consiste à étendre les alignements du jeu initial (les ancres initiales) grâce à la classification des données au sein des SOC. L'expansion de l'alignement est réalisée de manière récursive par comparaison des termes parents de termes t_1 et t_2, enfants de t_1 et t_2 et frères de t_1 et t_2.

Paramètres de filtre des alignements

Nous utilisons un filtre pour étudier les alignements calculés à partir des similarités syntaxiques. Le filtre `BestForBoth` que nous proposons permet de sélectionner les alignements qui sont les meilleurs à la fois pour le premier et le second terme.

3 Résultats et discussion

Évaluation des métriques de similarité syntaxiques

TABLE 1 *Résultat de l'évaluation des métriques de similarité syntaxique.*

Méthode	Filtre	Nombre d'alignements	Précision	Rappel
DL	Non	4 589	0,26	0,54
	`BestForBoth`	1 281	0,77	0,45
Stoilos (p_{diff} = 0.6 ; $p_{winkler}$ = 0.1)	Non	3 132	0,50	0,72
	`BestForBoth`	**1 356**	0,92	0,57
WGram	`BestForBoth`	2 322	0,63	0,67
WGram et Stoilos	`BestForBoth pour la métrique Stoilos`	1 202	**0,95**	0,53

On observe que le filtre `BestForBoth` augmente de manière significative (+0,5) la précision des alignement obtenus par méthode DL et Stoilos. Les meilleurs résultats sont obtenus par l'intersection des alignements de WGram et ceux de Stoilos. On observe que cette intersection permet de retrouver 50% des vrais alignements avec une précision de 95%.

Évaluation de la stratégie des ancres

Pour générer des alignements avec la méthode des ancres nous avons utilisé comme jeu de données initial les alignements créés par la collaboration entre le Regenstrief et l'IHTSDO. Le calcul des nouvelles ancres par la méthode DL en appliquant un seuil de 0,8 pour filtrer les alignements générés. On observe (tableau 2) que la combinaison des informations sémantiques et syntaxique permet d'augmenter de manière significative la précision (0,82). L'application des paramètres de filtre sémantique sur les alignements directement issus de la collaboration ont un rappel (0,37). Nous avons observé que 45% des alignements initiaux sont

composés par des termes LOINC® ou SNOMED CT ® dont aucune correspondance (CUI) n'a été retrouvée dans le Métathesaurus®.

TABLE 2 *Résultats de l'évaluation de la méthode des ancres. (1) La précision est calculée pour les alignements générés par la méthode des ancres grâce à une curation manuelle. (2) Le rappel des données filtrées a posteriori est calculé sur la base des vrais alignements.*

Méthodes	**Nombre d'alignements**	**Paramètres de filtres**	**Précision (1)**	**Rappel (2)**
Ancres DL à 0,80	1 833 ancres générés	NA	0,33	NA
		`BestForBoth`	0,58	**0,97**
		$(sim_{Semantique} = 1 \wedge confiance_{Semantique} > 800) \vee sim_{DL} = 1$	**0,82**	0,87
	2 177 initiales	$(sim_{Semantique} = 1 \wedge confiance_{Semantique} > 800)$	NA	**0,37**

4 Conclusion

Cette étude nous a permis d'identifier les forces et faiblesses des métriques et algorithmes étudiés. Nous avons notamment observé une amélioration de la précision par l'utilisation du filtre `BestForBoth` sur les métriques syntaxiques. L'algorithme heuristique des ancres filtré *a posteriori* par une dimension sémantique permet d'améliorer la précision des alignements obtenus. Cependant l'utilisation du Metathesaurus ne garantit pas de manière systématique (i) l'identification de concepts sémantiques pour l'ensemble des termes LOINC® et SNOMED CT® (ii) la résolution d'alignements entre termes complets et abrégés. Par la suite nous envisageons d'intégrer directement la similarité sémantique dans le processus de sélection des ancres. Pour résoudre les problèmes d'alignement lié aux termes abrégés nous envisageons de construire ou de réutiliser un dictionnaire d'abréviation, spécifique au domaine du diagnostic *in vitro*.

Référence

DAMERAU, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, *7*(3), 171–176.

EUZENAT, J., & SHVAIKO, P. (2013). *Ontology matching*. Springer-Verlag.

FUNG, K. W., & BODENREIDER, O. (2005). Utilizing the UMLS for Semantic Mapping between Terminologies. *AMIA Annual Symposium Proceedings*, *2005*, 266‑270.

HODGE, G. (2000). *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. ERIC.

IHTSDO & REGENSTRIEF INSTITUTE. (juillet 2013). Regenstrief and the IHTSDO are working together to link LOINC and SNOMED CT. Repéré à https://loinc.org/collaboration/ihtsdo

LEVENSHTEIN, V. I. (1966). *Binary codes capable of correcting deletions, insertions, and reversals*. Communication présentée au Soviet physics doklady (vol. 10, p. 707–710).

MACARY, F. (2007). IHDE, CDA et LOINC : des composants d'interopérabilité au service du partage des résultats de biologie médicale. *Spectra biologie*, *26*(158), 51‑57.

SEDDIQUI, M. H., & AONO, M. (2009). An efficient and scalable algorithm for segmented alignment of ontologies of arbitrary size. *Web Semantics: Science, Services and Agents on the World Wide Web*, *7*(4), 344‑356. doi:10.1016/j.websem.2009.09.001

STOILOS, G., STAMOU, G., & KOLLIAS, S. (2005). A String Metric for Ontology Alignment. Dans Y. Gil, E. Motta, V. R. Benjamins, & M. A. Musen (dir.), *The Semantic Web – ISWC 2005* (p. 624‑637). Springer Berlin Heidelberg.

Détection et Représentation des changements dans les sources de données RDF

Daniel Mercier[1], Nathalie Pernelle[1], Fatiha Saïs[1], Sujeeban Thuraisamy[1]

UNIVERSITÉ PARIS SUD, LABORATOIRE DE RECHERCHE EN INFORMATIQUE
91405 Orsay cedex, France
{pernelle, sais}@lri.fr, {DanielMercier,Sujeeban.Thuraisamy}@u-psud.fr

Résumé : De nombreuses sources de données RDF sont en évolution constante que ce soit au niveau des données ou du vocabulaire utilisé. Or, de nombreuses tâches d'intégration sont impactées par ces modifications. Nous présentons une approche permettant de détecter et de représenter des changements plus ou moins complexes que l'on peut détecter lorsque l'on s'intéresse aux seules données. Une première expérimentation a été menée sur différentes versions de DBPedia.

Mots-clés : Ontologies, Représentation des connaissances, Evolution des données

1 Introduction

Le Web des données est en évolution permanente. De nombreuses modifications sont apportées quotidiennement sur les données et les vocabulaires publiés sur le LOD. En 2012, 76% des documents RDF du LOD ont subi au moins une modification (Käfer *et al.* (2012)). Ces modifications peuvent intervenir au niveau ontologique (classes, propriétés, axiomes et correspondances avec les autres ontologies) ou/et au niveau données (entités, valeurs des propriétés et liens d'identité entre entités). De nombreuses tâches d'intégration de données sont impactées par ces modifications (e.g synchronisaton, liage, fusion de données). Dans ce contexte, il est important de disposer d'outils permettant de détecter et de représenter ces changements de façon à ce que les tâches impactées puissent mettre à jour leurs résultats sans devoir être ré-exécutées sur toutes les données. De nombreux travaux se sont focalisés sur la détection, la représentation et la gestion des changements au niveau ontologique (Zablith *et al.* (2015); Dinh *et al.* (2014)). Quelques travaux récents (Papavasileiou *et al.* (2013); Roussakis *et al.* (2015)) se sont intéressés au problème de détection de changement dans les données. Cependant, ces derniers ne sont pas représentés dans une ontologie ou sont synthétiques et ne gardent pas trace des triplets impliqués. Nous proposons une approche permettant de détecter et de représenter sémantiquement des changements plus ou moins complexes que l'on peut identifier lorsque l'on s'intéresse aux seules données.

2 Approche de détection et de représentation de l'évolution de sources de données RDF

Etant données deux versions TDB_{old} et TDB_{new} d'une source de données, nous détectons tout d'abord les ensembles de de triplets ajoutés et supprimés. Ensuite, une étape de représentation sémantique des changements est réalisée. Pour cela, nous avons défini une ontologie appelée O^{DE} qui permet de représenter sémantiquement les types de changements. Deux types généraux de changements sont distingués *AddedStatement* et *DeletedStatement* pour lesquels

quatre sous-types sont définis. Pour le type *AddedStatement* nous avons défini [1] :
-*AddedSchemaElement* qui représente les assertions impliquées dans l'introduction d'une nouvelle propriété (sous-type *AddedProperty*) ou d'une nouvelle classe (sous-type *AddedClass*).
-*AddedTyping* qui regroupe les assertions concernant l'introduction de nouveaux types pour des instances existantes. Parmi ces assertions, certaines concernent la première utilisation d'une classe dans la source de données (sous-type*AddedClass*).
-*AddedInstance* qui représente l'ensemble des assertions décrivant des instances nouvelles.
InstanceDescriptionEnrichment qui décrit l'ensemble des assertions qui viennent enrichir des instances existantes.
Chaque triplet supprimé ou ajouté est réifié et automatiquement associé à un ou plusieurs types de changements définis dans O^{DE}. L'ontologie peuplée peut ensuite être exploitée pour répondre à des requêtes d'experts simples (e.g. le nombre d'instances supprimées) ou plus complexes (e.g. les propriétés fonctionnelles dont la valeur a été modifiée).

TABLE 1 – Requêtes simples exploitant la classe DeletedClass

/* Liste des classes supprimées

```
SELECT DISTINCT ?deletedClass
WHERE {
?node rdf:type   ode:DeletedClass
?node rdf:object ?deletedClass .
}
```

/* Instances conduisant à une suppression de classe

```
SELECT ?s
WHERE {
?node rdf:type ode:DeletedClass .
?node rdf:subject ?s .
}
```

3 Premières experimentations

Nous avons évalué notre approche en utilisant les versions 3.5, 3.8 et 3.9 de DBPedia [2]. Nous nous sommes intéressés à la classe $Person$ (fichier PersonData). Le tableau 2 présente le nombre de triplets, le nombre d'instances, le nombre de propriétés ainsi que le nombre de types différents pour les trois différentes versions de cette classe.

TABLE 2 – Evolution des triplets décrivant les Personnes dans trois versions de DBPedia

	Version 3.5	Version 3.8	Version 3.9
#triplets	482 080	18 719 429	22 008 122
#instances	48 692	2 853 529	3 733 629
#propriétés	9	9	9
#types	71	348	434

Nous avons détecté les changements élémentaires entre deux versions successives de la classe $Person$ et nous avons peuplé l'ontologie O_{DE} (temps d'exécution sur plus de 18 millions de triplets : 55 mn, 155 millions de triplets après réification). Presque toutes ces assertions sont de type $AddedStatement$ (tab. 3), mais près de la moitié des triplets existants ont été supprimés

1. Les sous-classes de *DeletedStatement* peuvent être décrites de façon symétrique.
2. http ://dbpedia.org/services-resources/datasets

(classe $DeletedStatement$). La table 3 montre par exemple comment les assertions instances de la classe $AddedStatement$ se répartissent dans les différentes classes de l'ontologie O_{DE}.

TABLE 3 – Type et nombre de changements pour la classe $AddedStatement$

	#Added Statements	#Added SchemaElement	#Added Property	#Added Class	#Added Typing	#Added Instance
v3.5 -> v3.8	18 469 394 (12 :43 mn)	284 (5 :16 mn)	5 (3 :28 mn)	279 (1 :48 mn)	13 596 447 (6 :43 mn)	2 835 666 (7 :20 mn)
v3.8 ->v3.9	4 813 958 (01 :49 mn)	86 (14 s)	0 (2 s)	86 (12 s)	4 015 870 (1 :21 mn)	1 103 520 (45 s)

Les classes $addedTyping$ et $deletedtyping$ montrent que le typage des instances évolue mais l'approche permet également de détecter que les éléments de l'ontologie utilisés ont évolué. Ainsi, entre v3.5 et v3.8, 284 classes sont apparues dans la description des personnes et deux ont disparu. Nous avons exécuté différentes requêtes permettant par exemple d'utiliser O_{DE} pour lister les nouvelles instances d'une sous-classe donnée ou pour rechercher les assertions de type $InstanceDescriptionEnrichment$ qui ont été ajoutées pour une URI donnée. Une requête simple permet ainsi de montrer que 7 assertions ont été ajoutées pour Barak Obama (e.g. la propriété $description$ prend la valeur "American politician, 44th President of the United States"@en dans v3.8) et que cette URI est nouvellement typée par les classes $Agent$ et $OfficeHolder$.

4 Conclusion

Nous proposons une approche originale de détection et de représentation sémantique des changements dans une source RDF. Nous avons conçu une ontologie O^{DE} qui représente sémantiquement les différents types de changements pouvant survenir au niveau des données et qui est automatiquement peuplée par notre approche. Cela permet à un expert ou à un outil d'interroger l'ontologie via des requêtes SPARQL pour observer l'évolution des données. Nous souhaitons maintenant détecter d'autre types de changements (changements induits par des sources externes via des liens $SameAs$, axiomes).

Références

DINH D., DOS REIS J. C., PRUSKI C., SILVEIRA M. D. & REYNAUD-DELAÎTRE C. (2014). Identifying relevant concept attributes to support mapping maintenance under ontology evolution. *J. Web Sem.*, **29**, 53–66.

KÄFER T., UMBRICH J., HOGAN A. & POLLERES A. (2012). Dyldo : Towards a dynamic linked data observatory. In *WWW2012 Workshop on Linked Data on the Web, Lyon, France, 16 April, 2012*.

PAPAVASILEIOU V., FLOURIS G., FUNDULAKI I., KOTZINOS D. & CHRISTOPHIDES V. (2013). High-level change detection in rdf(s) kbs. *ACM Trans. Database Syst.*, **38**(1), 1 :1–1 :42.

ROUSSAKIS Y., CHRYSAKIS I., STEFANIDIS K., FLOURIS G. & STAVRAKAS Y. (2015). A flexible framework for understanding the dynamics of evolving RDF datasets. In *The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA*, p. 495–512.

ZABLITH F., ANTONIOU G., D'AQUIN M., FLOURIS G., KONDYLAKIS H., MOTTA E., PLEXOUSAKIS D. & SABOU M. (2015). Ontology evolution : a process-centric survey. *Knowledge Eng. Review*, **30**(1), 45–75.

Représentation sémantique des documents pédagogiques

Nawel SEKKAL[1], Fedoua DIDI[2]

1 n_iles@mail.univ-tlemcen.dz
LRIT - Laboratoire de Recherche en Informatique de Tlemcen
Université Abou Bekr Belkaid Tlemcen , Algérie
2 fedouadidi@yahoo.fr
LRIT - Laboratoire de Recherche en Informatique de Tlemcen
Université Abou Bekr Belkaid Tlemcen , Algérie

Résumé : Une masse importante de ressources pédagogiques est produite à travers de nombreuses universités, et cela est dû à l'évolution rapide des moyens d'information et de communication dans le domaine de l'apprentissage automatique. Nous proposons un modèle d'entrepôt qui vise à pallier ce déficit. Il s'agit d'un modèle d'entrepôt distribué qui repose sur la description par les métadonnées de LOM et l'indexation sémantique des objets pédagogiques. Nous supposons dans notre approche que chaque université dispose de son propre entrepôt, contenant des objets pédagogiques, leur description en métadonnées et leur description sémantique. Les métadonnées et les descripteurs de chaque entrepôt sont alors utilisés pour capitaliser les connaissances sur les objets distribués et alimenter un méta entrepôt, comparable à un catalogue accessible par tous.

Mots-clés : apprentissage automatique, objet pédagogique, entrepôt, métadonnée LOM, indexation sémantique, partage, réutilisation.

1. Introduction et problématique

Les Technologies de l'Information et de la Communication (TIC) améliorent profondément nos façons de nous informer, de communiquer et de nous former. Dans le domaine universitaire, les objets pédagogiques « OP » constituent l'ensemble des informations, parties de cours, de programmes, de thèses, etc. qui permettent de véhiculer et de transmettre des concepts et contenus d'enseignements. Compte tenu du coût de production de tels OP et de l'expertise nécessaire pour les produire, il est primordial de les rendre facilement accessibles, exploitables et réutilisables. Notre proposition vise à mettre à disposition les OP existants à un ensemble d'universités organisées en réseau et de garantir leur réutilisation. À l'heure du Web sémantique, il est alors nécessaire d'associer des annotations sémantiques au contenu des OP pour permettre de les retrouver. Nous proposons dans ce travail de recherche un modèle d'entrepôt distribué qui repose sur la description des OP par les métadonnées de LOM et les concepts des ontologies de domaine.

2 . Représentation et modélisation de l'entrepôt pédagogique distribué

2.1. Architecture d'un entrepôt local

Le contenu de l'entrepôt pédagogique associé à une université, doit être constitué d'un ensemble de ressources pédagogiques de différents formats (pdf, doc, jpg, etc.) et sur différents supports (papier, électronique, multimédia, etc.). Une ressource pédagogique doit être sélectionnée et filtrée avant d'être stockée dans l'entrepôt. Elle est composée d'un ensemble d'objets pédagogiques reliés entre eux. Chaque objet est décrit par les métadonnées LOM afin de le retrouver et de le réutiliser. Notre entrepôt local est représenté par l'architecture suivante :

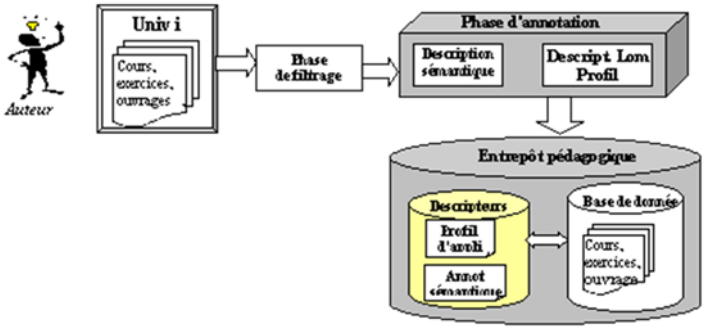


FIGURE 1. Modélisation d'un entrepôt local.

L'intégration d'un OP dans l'entrepôt distribué se fait en deux phases :
1- Phase de filtrage : il s'agit de sélectionner les OP pertinents pour les capitaliser, tout en gardant les relations entre les objets appartenant à la même ressource.

2- Phase de qualification : à chaque OP sont associés un ou des descripteurs LOM et une ou des annotations sémantiques par rapport aux ontologies de domaine (Snae, 2007). La mise en relation entre les ontologies et les contenus des OP s'appuie sur les techniques d'indexation automatique telles qu'abordées par Hernandez (2008).

- La description par les métadonnées de LOM permet de décrire et d'indexer tout OP. Les métadonnées obtenues permettent de donner les différents renseignements nécessaires sur chaque OP. Elles seront filtrées par un profil d'application. Cependant, comme nous avons déjà mentionné, la représentation sémantique se basant sur les normes n'est pas suffisante pour permettre la réutilisation des OP dans d'autres applications.

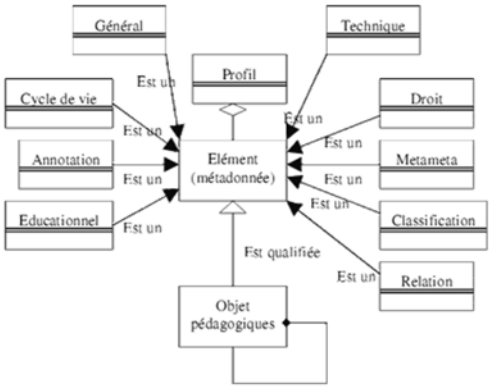


FIGURE 2. Description d'un objet pédagogique par LOM.

- Afin d'assurer la réutilisation d'un OP, la représentation sémantique est basée sur la notion d'ontologie de thème (Hernandez, 2008) de la formation qui rassemble les thèmes, notions et connaissances à appréhender (figure 3).

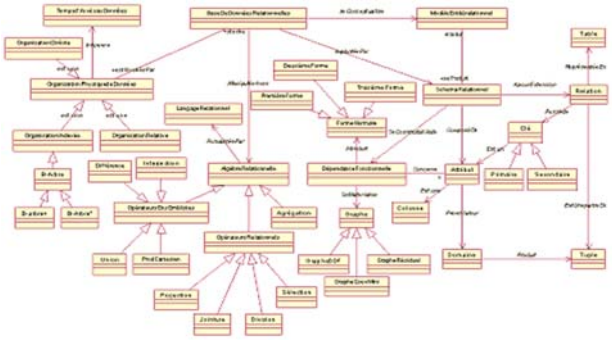


FIGURE 3. Extrait de l'ontologie du thème de l'informatique (Hernandez, 2008). [Zoom]

2.2. Architecture générale du méta entrepôt pédagogique

Dans le cadre universitaire, il existe différentes sources des OP : supports de cours, exercices, thèses, rapports, etc., disponibles par les utilisateurs au niveau de chaque université. Cependant, il leur est impossible d'accéder de façon efficace, et de trouver les OP pertinentes adaptés à leurs besoins. D'un autre coté, ces OP restent localement réservés, et non accessibles par les utilisateurs d'autres d'universités. Dans notre approche, nous proposons une architecture distribuée qui consiste à mettre en place un système qui va permettre de :

- Capitaliser au niveau de chaque université, tous les OP pertinents produites ;
- Annoter chaque OP afin de la rendre accessible et réutilisable. Cette annotation se base sur une description par les métadonnées LOM et par les annotations sémantiques ;
- Créer un méta-entrepôt pédagogique qui sera alimenté par l'ensemble des descripteurs des entrepôts des universités. Ainsi chaque ressource pourra être partagée entre universités et elle sera accessible à distance (tout en respectant les rassemblées dans un endroit unique.

L'architecture générale du modèle est la suivante :

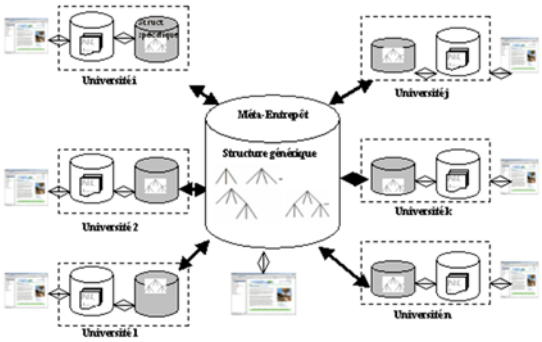


FIGURE 4. Architecture distribuée de l'entrepôt pédagogique.

Les entrepôts pédagogiques se basent sur deux types de structures, des structures spécifiques et une structure générique : La première caractérise l'organisation d'un OP où chaque entrepôt peut contenir plusieurs structures logiques spécifiques. Elle permet la description sémantique d'un OP. La deuxième représente une collection de représentation spécifique. Les entrepôts ayant des structures similaires sont regroupés dans une même structure spécifique afin de faciliter la recherche.

3. Conclusion

Pour conclure, nous avons proposé un modèle d'entrepôt distribué qui permet l'indexation des objets pédagogiques, leur recherche et leur réutilisation de façon à répondre aux besoins des apprenants ou des enseignants. L'originalité de notre approche peut être résumée par le fait qu'elle s'appuie sur un enrichissement des descriptions LOM par des annotations sémantiques par rapport aux ontologies de domaine ; elle repose sur une architecture distribuée où les spécificités de chaque université sont prises en compte ; permet une interrogation homogène des objets existants via une méta description des objets ; et enfin elle assure des possibilités avancées de traitement via l'utilisation des fonctionnalités d'un entrepôt de données.

Références

MIHAELA M. BRUT, FLORENCE SEDES, ET STEFAN DANIEL DUMITRESCU, (2011), A SEMANTIC-ORIENTED APPROACH FOR ORGANIZING AND DEVELOPING ANNOTATION FOR E-LEARNING. IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, VOL. 4, NO. 3 .

HERNANDEZ, N., MOTHE, J., RALALASON, B., RAMAMONJISOA, B., STOLF, P. (2008). A Model to Represent the Facets of Learning Objects, Santa Rosa-USA.

Snae, C., Brueckner, M. (2007). Ontology-Driven E-Learning System Based on Roles and Activities for The Learning Environment, Volume 3, p. 1-17.

Index des auteurs

Sponsors

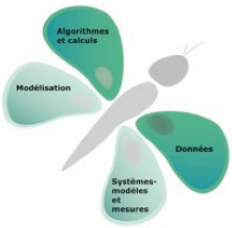

AfIA
Association française
pour l'Intelligence Artificielle


Institut de biologie
computationnelle


LANGUEDOC
ROUSSILLON
LA RÉGION MIDI
PYRÉNÉES


cnrs


LIRMM


Algorithmes
et calculs
Modélisation
Données
Systèmes-
modèles
et
mesures


MIASHS


SIFR
project


WEARELEARNING


WEARELEARNING


expernova


irstea


UNIVERSITÉ
PARIS
SUD
IUT DE SCEAUX
université PARIS-SACLAY