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.

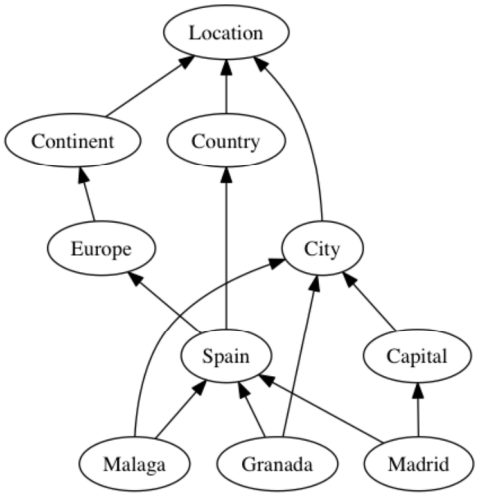

FIGURE 1 – Exemple d'un ordre partiel entre certaines valeurs, qui inclue les relations de spécialisation **`subClassOf`** et de composition **`partOf`**

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.

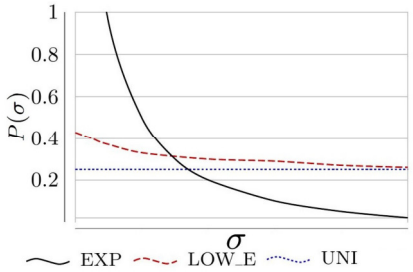

FIGURE 2 – Distributions utilisées pour la sélection des valeurs "vraies"

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**

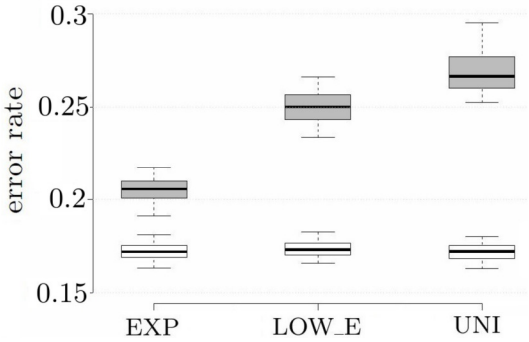

FIGURE 3 – Taux d'erreur en fonction des différents types de test : *Sums* en gris, *Adapted_Sums* en blanc

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)