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-

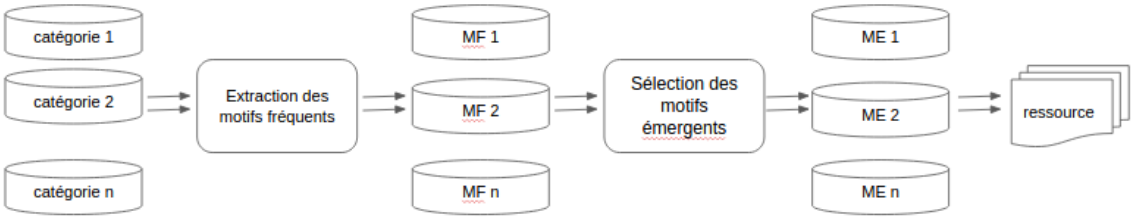

FIGURE 1 – Vue générale du processus de création de ressource

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 reférence (avec et sans enrichissement) et après comparer les groupes prédits et les groupes de reférence en utilisant les mesures décrites dans la section 3.2.

3.1 Données de reférence

Les données de reférence 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 reférence. 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.