SHMC - Semantic Hierarchical Multi-label Classification

Approche Big Data et Web Sémantique pour la classification automatique de données web et la recommandation d'articles économiques

Christophe Cruz[1]
Christophe.cruz@ubfc.fr

[1] IUT de Dijon-Auxerre, Univ. Bourgogne – Franche-Comté
Laboratoire Le2i, FRE 2005 CNRS

Thèmes – *Informatique - Économie*

Résumé – *La veille stratégique regroupe plusieurs activités parmi lesquelles l'analyse, la découverte, le croisement et la synthèse de l'information sont des tâches de haut niveau effectuées par des experts dans leur domaine. Dans un contexte web, le volume grandissant des données, leur grande variété et la vitesse à laquelle les données évoluent, font apparaître un besoin d'automatiser tout ou partie de ces tâches. Créer des outils de résolution automatique de ces tâches demande d'une part la mise en place d'un processus d'apprentissage automatique permettant d'imiter des processus de décision complexes, et d'autre part la prise en compte des connaissances métier des experts dans le processus d'apprentissage. Cet article décrit une proposition d'architecture et de méthodologie pour l'indexation automatique d'article économique afin de les recommander aux clients de l'entreprise Actualis Sarl, partenaire du projet.*

Mots-Clés – *Système de recommandation, apprentissage artificiel, ontologie, modélisation des connaissances*

1 Introduction

Afin de rester en phase avec les tendances actuelles du marché, le processus de prise de décision dans le domaine économique nécessite la centralisation et l'apport de grandes quantités d'informations. Pour cela, les hommes d'affaires, les entrepreneurs et les vendeurs doivent parfaitement connaître leur environnement. Cela signifie qu'il faut maintenir une veille économique constante facilitant l'identification des perspectives d'affaires, permettant de décrocher de nouveaux contrats.

L'entreprise Actualis SARL (http://firsteco.fr/) partenaire de ce projet est spécialisée dans la production et la distribution de revues de presse économiques. Le web est par définition une masse de données immense, et une source d'information clé dans le domaine de la veille économique. Ainsi, de nombreuses entreprises sous-traitent ce processus de veille économique et stratégique à des sociétés spécialisées dans le domaine.

Dans ce contexte, nous avons proposé une méthode pour enrichir sémantiquement une ontologie utilisée pour classer des articles de presse, en utilisant un raisonneur sémantique basé sur la logique de description (DL) [1]. Nous avons également proposé une architecture pour extraire des informations de valeur à partir de larges volumes de données textuelles dans un contexte Big Data, en utilisant un processus de Classification Hiérarchique Multi-étiquette Sémantique [2, 3, 4]. SHMC est un processus d'apprentissage d'ontologie non supervisé, basé sur les technologies du big data, de l'apprentissage artificiel, et du raisonnement logique à base de règles. Cette approche offre la possibilité de construire un système de recommandation hybride exploitant à la fois les profils des utilisateurs créer à partir des vocabulaires contrôlés et des articles de presse indexés à l'aide des mêmes vocabulaires contrôlés. Les articles et les connaissances associées sont stockés dans la base de connaissances formant ainsi le point central de l'architecture (cf. figure 1.).

En 2001, un rapport de recherche du groupe Gartner décrivait pour la première fois la notion de « Big Data » et les caractéristiques qui définissent ce domaine. Dans ce rapport, la problématique de gestion de ce nouveau type de données est déjà définie autour de trois dimensions, le Volume, la Vélocité et la Variété de ce type de données. Le Volume définit la quantité croissante de données, générée et stockée au fil du temps par les réseaux sociaux, les données de capteurs, d'objets connectés, etc. [5]. La

Vélocité concerne la vitesse importante de production des données, et par conséquent le besoin de traitement rapide des données. La Variété représente la grande hétérogénéité des formats du Big Data. En particulier, les données non-structurées et semi-structurées nécessitent des traitements importants, et représentent 90% du contenu associé généralement au Big Data (pages web, documents en langage naturel, audio, etc.) [6].
D'autres dimensions nommées par des "V" supplémentaires ont émergé au cours du temps, comme la Véracité, la Visualisation ou la Valeur par exemple. Contrairement aux 3V du Big Data, il ne s'agit plus d'indicateurs quantitatifs, liés à la robustesse face aux données, et ne sont donc pas considérées comme des dimensions au même titre que le Volume, la Vélocité et la Variété. En revanche, la Valeur définit la pertinence d'une information pour l'utilisateur final, et cette caractéristique est centrale dans ce projet.

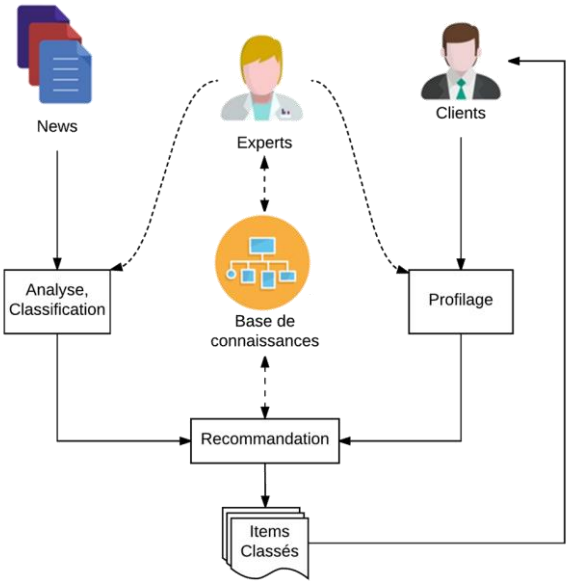

Figure 1 – Architecture SHMC

A noter que les entreprises utilisent de façon récurrente des taxonomies de domaine, afin de représenter leur connaissance métier. Ils formalisent ainsi un modèle associant une valeur aux données [7]. Dans le cadre de la veille économique, ce type de taxonomie permet donc de qualifier une information de façon précise. Cette qualification, qui se base sur les concepts clés du domaine de l'entreprise est un vecteur de valeur pour un client.
Dans la suite, nous présenterons le positionnement des travaux, ainsi que l'apport méthodologique du projet.

2 Positionnement

L'utilisation des ontologies pour la tâche de classification porte souvent sur la description du modèle de classification (étiquette, items, règles de classification). Galinina et Borisov [8] utilisent deux ontologies dans un système de classification : (1) une ontologie de domaine, indépendante de la méthode de classification, et (2) une ontologie dédiée à la méthode de classification basée sur un arbre de décision. Au-delà de la description du domaine, les ontologies sont utilisées pour améliorer le processus de classification. Elberrichi et al. [9] présente une méthode en deux étapes pour améliorer la classification de documents médicaux (MeSH - Medical Subject Headings). Leurs résultats montrent que l'utilisation d'une ontologie de domaine permet d'améliorer la performance de la méthode de classification de documents.
La plupart des travaux de la littérature se concentrent sur l'amélioration du processus de classification en utilisant les ontologies, ce qui permet d'améliorer la description des items. En revanche, ils ne tirent pas avantage des capacités des raisonneurs sémantiques pour classer automatiquement des items [2].
De plus, les ontologies permettent d'améliorer le processus de recommandation grâce à la description sémantique des articles. Deux principaux systèmes de recommandation sont distingués, les systèmes dits de filtrage collaboratif et les systèmes basés sur le contenu. Dans le cas présent, nous nous focalisons sur la modélisation sémantique des caractéristiques du contenu des articles par une approche ontologique. Cette approche nous permet de définir un ensemble de vocabulaires contrôlés hiérarchiques du domaine de métier qui caractérisent chacune des facettes de l'article. Ainsi, chaque facette permettra de définir une dimension descriptive de l'article sous la forme de vecteur [1].

3 Apport méthodologique

Le processus SHMC permettant d'indexer les articles économiques se compose de 5 étapes. Ces étapes correspondent à l'élément « Analyse, Classification » de la figure 1.

- **Indexation** : extrais les termes des items (documents textes), et crée un index inversé des items.
- **Vectorisation** : calcule les vecteurs de fréquence des termes à partir de l'index inversé. L'ensemble de vecteurs de termes permet de générer une matrice de cooccurrences des termes.
- **Hiérarchisation** : détermine les termes les plus pertinents, i.e. les Labels, et génère une hiérarchie de subsomption des labels à partir de la matrice.
- **Résolution** : créé des règles de classification qui lient les nouveaux items aux labels à partir de la matrice de cooccurrence.
- **Réalisation** : remplis l'ontologie avec les nouveaux items et pour chaque item détermine les labels les plus spécifiques.

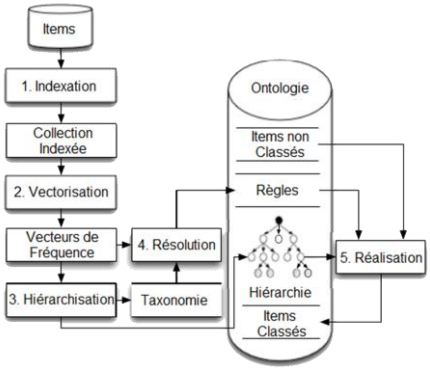

Figure 2 – Processus SHMC

La figure 2 présente la séquence des étapes du processus SHMC. La figure suivante présente les différentes étapes du traitement automatique du langage appliqué sur les articles. Ces étapes sont réalisées dans la première phase correspondant à la phase « indexation » du processus SHMC.

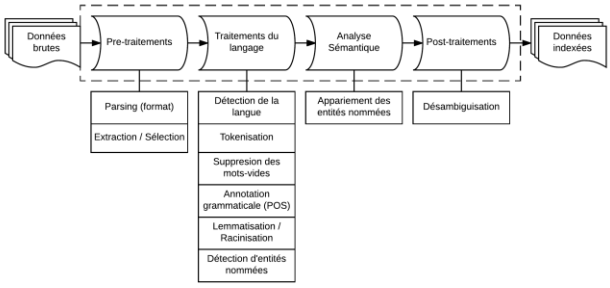

Figure 3 – Indexation : traitement automatique du langage

Le processus SHMC est un processus Big Data, car ils manipulent un grand volume de données impliquant le développement d'algorithmes respectant le paradigme Map Reduce pour chaque processus de SHMC. Ainsi, ces algorithmes se caractérisent par leur capacité à passer l'échelle. Par exemple, la figure 4 présente la matrice de fréquences ou matrice de cooccurrences des termes dans les documents. La matrice est composée de 10^6 mots ou dimension (features) et de 10^4 étiquettes (labels). Cela représente une matrice de 10^{10} décimaux. La matrice résultante ne peut pas être stockée dans un seul nœud de serveur et doit donc être répartie à l'aide d'une base de données NoSQL qui est en l'occurrence HBase. L'ensemble l'implémentation est réalisée à l'aide du framework Hortonworks (https://hortonworks.com/).

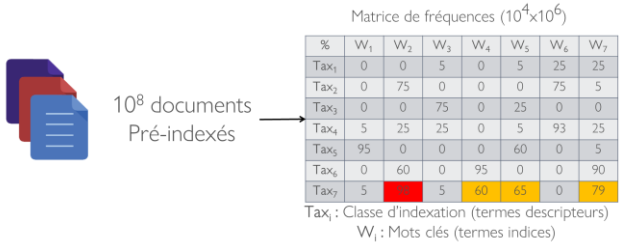

Figure 4 – Vectorisation : matrice de fréquences

4 Conclusion

Cet article a décrit un processus de classification hiérarchique multi-étiquette nommé SHMC traitant des données textuelles non structurées dans un contexte Big Data. Les résultats obtenus, mais non présentés ici, montrent que l'approche basée sur l'apprentissage automatique d'une ontologie et le raisonnement sémantique revêt un caractère très intéressant par sa dimension explicative du résultat d'indexation. Cette caractéristique ne peut pas être obtenue par d'autres méthodes telles que les réseaux de neurones.

5 Remerciements

Les auteurs souhaitent remercier les financeurs des travaux sur ce projet, à savoir l'entreprise Actualis SARL, l'agence française ANRT, le conseil régional de Bourgogne Franche-comté et le Fonds Européen de Développement Economique et Régional (FEDER).

Références

[1] D. Werner, N. Silva, C. Cruz, A. Bertaux, *Using DL-reasoner for hierarchical multilabel classification applied to economical e-news*. In Proceedings of 2014 Science and Information Conference, SAI 2014, pp. 313–320.

[2] R. Peixoto, T. Hassan, C. Cruz, A. Bertaux, N. Silva, *Semantic HMC: A Predictive Model using Multi-Label Classification For Big Data*. In The 9th IEEE International Conference on Big Data Science and Engineering (IEEE BigDataSE-15), 2015

[3] R. Peixoto, T. Hassan, C. Cruz, A. Bertaux, N. Silva, *An unsupervised classification process for large datasets using web reasoning*. In ACM (Ed.), SBD'16: Semantic Big Data Proceedings, San Francisco (CA), USA, 2016

[4] R. Peixoto, T. Hassan, C. Cruz, A. Bertaux, N. Silva, *Hierarchical multi-label classification using web reasoning for large datasets*. Open Journal of Semantic Web (OJSW) 3(1), 1–15, 2016

[5] M. Chen, S. Mao, Y. Liu, *Big data: A survey. Mobile Networks and Applications*, 19(2) :171–209, 2014

[6] A. R. Syed, K. Gillela, C. Venugopal, *The Future Revolution on Big Data*. International Journal of Advanced Research in Computer and Communication Engineering, 2(6) :2446–2451, 2013

[7] P. Lambe, *Organising knowledge: taxonomies, knowledge and organisational effectiveness*. Elsevier Science, ISBN-13: 978-1843342274, 2014

[8] A. Galinina, A. Borisov, *Knowledge modelling for ontology-based multiattribute classification system*. Applied Information and Communication, 103–109, 2013

[9] Z. Elberrichi, B. Amel, T. Malik. *Medical Documents Classification Based on the Domain Ontology MeSH*. arXiv preprint arXiv :1207.0446., 2012