Classification multi-label par raisonnement logique pour l'indexation sémantique de documents

David Werner*, Christophe Cruz* et Aurelie Bertaux*

*Université de Bourgogne, LE2I
david.werner@u-bourgogne.fr
christophe.cruz@u-bourgogne.fr
aurélie.bertaux@iut-dijon.u-bourgogne.fr

Résumé. Cet article présente une solution centrée sur les ontologies pour la classification multi-label automatique d'information nécessaire à un système de recommandation d'informations économiques.

1 Introduction

Les systèmes de recommandation basés sur le contenu suivent généralement un processus en deux étapes : (i) Création d'une représentation du besoin des utilisateurs ainsi que des informations à recommander. (ii) Comparaison des représentations afin d'évaluer la pertinence d'une information pour un utilisateur en fonction de son profil. Notre approche consiste à automatiser l'indexation à l'aide de processus d'inférence sur une ontologie d'indexation intégrant les vocabulaires contrôlés (e.g. thésaurus, nomenclatures, listes) définis par les documentalistes pour modéliser le domaine. Le respect de la vision métier sur le domaine permet une supervision simplifiée pour les documentalistes, garantissant la qualité de l'indexation.

2 Automatisation du processus d'indexation

La classification multi-label consiste à associer des étiquettes à des items (Tsoumakas et Katakis, 2007). Cet article propose une méthode pour enrichir sémantiquement une ontologie en adoptant des processus d'apprentissage automatique pour indexer et décrire l'indexation de façon à réduire l'écart entre le point de vue des experts et les règles d'indexation. L'approche proposée repose sur les quatre phases suivantes :

Phase 1 : utilisation du travail d'indexation déjà fait par les documentalistes et d'un processus d'analyse de texte pour extraire des mots-clés afin de générer une matrice qui présente la fréquence de chaque mot-clé en fonction de chaque étiquette.

Phase 2 : utilisation de la matrice afin de définir des règles capables de déterminer si un document doit être associé à une étiquette sur la base des mots-clés qu'il contient. Deux seuils de fréquence sont définis, α et β. Les mots-clés dont la fréquence est supérieure au seuil α sont considérés comme des indices fiables. La présence d'un seul de ces mots est considérée comme suffisante pour que le document soit associé à l'étiquette. Le seuil de fréquence inférieur est β. Dans ce cas, nous avons besoin d'une combinaison de β-termes (dont la fréquence

est supérieure à β) pour prendre la décision d'associer un document avec l'étiquette. Plus d'informations sur les règles d'indexation peuvent être trouvées dans (Werner et al., 2014).

Phase 3 : la classification fournit deux types de résultats. Le premier est la découverte de la classe de subsomption la plus spécifique. Le second est la déduction des classes d'équivalence lorsque les contraintes logiques sont équivalentes. D'une part, cela signifie que lorsqu'un document est étiqueté (lors de la phase 4) par une classe qui possède des subsumants, ce document est également marqué par les classes subsumantes. D'autre part, lorsqu'un document est étiqueté avec une classe qui a des classes d'équivalence alors ce document est également étiqueté avec ces classes équivalentes. Ces deux éléments peuvent permettre la classification multi-label. De plus, sachant que les étiquettes peuvent être organisées de façon hiérarchique il peut s'agir d'un processus de classification hiérarchique multi-label (HMC).

Phase 4 : la phase de réalisation consiste à trouver toutes les classes les plus spécifiques des individus. Cette phase est mise en œuvre par le moteur d'inférence. Les phases 3 et 4 utilisent des raisonneurs comme FaCT ++, HermiT ou Pellet.

3 Conclusion

Cet article aborde l'indexation de documents de façon automatique sur la base de vocabulaires contrôlés contenus dans une ontologie OWL-DL à l'aide de raisonneurs. Nous décrivons le processus, qui consiste à enrichir une ontologie existante afin de l'utiliser pour le processus de classification multi-label automatique d'articles de presse économique. Nos tests préliminaires à l'aide du jeu de données *del.icio.us* disponible sur le site web du projet Mulan [1] ont mis en évidence la complexité du raisonnement sur l'ontologie. L'utilisation de règles simples donne de faibles résultats, alors que l'utilisation de règles plus complexes pouvant donner de meilleurs résultats impacte de façon très importante les performances (charge mémoire et temps de calcul). Nos futurs travaux visent à effectuer les raisonnements à l'aide de règles complexes sur des sous-parties de l'ontologie, l'ontologie entière étant trop lourde.

Références

Tsoumakas, G. et I. Katakis (2007). Multi-label classification : An overview. *Int J Data Warehousing and Mining 2007*, 1–13.

Werner, D., N. Silva, C. Cruz, et A. Bertaux (2014). Using dl-reasoner for hierarchical multilabel classification applied to economical e-news. In *Science and Information Conference (SAI), 2014*, pp. 313–320.

Summary

This article presents the automatic multi-label classification task of information for a recommender system of economic news.

1. http ://mulan.sourceforge.net/