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