Les règles d'association comme outil de catégorisation textuelle

Simon Jaillet, Maguelonne Teisseire, Jacques Chauché

LIRMM - CNRS - Université Montpellier 2
161 rue Ada, 34392 Montpellier Cedex 5 France
{jaillet, teisseire, chauche}@lirmm.fr

Résumé

Nous proposons d'étudier les performances d'une méthode de catégorisation automatique de textes basée sur des règles d'association avec deux autres méthodes de classification (SVM et Rocchio). Ce comparatif est réalisé non pas sur une représentation textuelle classique, mais sur une représentation des textes basée sur des vecteurs conceptuels. En effet, la représentation textuelle la plus utilisée est du type $TF\text{-}IDF$ (SMART) mais dans [Jaillet *et al.*, 2003], les auteurs proposent une nouvelle méthode de représentation des documents. Dans cette nouvelle approche, au lieu de définir un espace vectoriel dont chaque dimension représente un terme d'indexation souvent assimilé à un stem (radical), l'ensemble des termes est projeté sur un ensemble fini de concepts extrait d'un thesaurus. L'approche de catégorisation proposée est une amélioration de CBA [Liu *et al.*, 1998]. La première modification concerne le filtrage des items fréquents par une valeur d'entropie. Ce pré-traitement est utile afin de limiter les effets exponentiels de l'algorithme APRIORI lors de la génération des candidats. La deuxième modification concerne la gestion du support. Pour CBA, le support minimum à vérifier est fixé pour toute la base. Or, dans la majorité des cas, les catégories ne possèdent pas la même fréquence de distribution. Nous proposons de résoudre ce problème en divisant le jeu d'entraînement en n sous-bases, n représentant le nombre de catégories du corpus. Ensuite, la recherche de fréquents s'effectue sur chacune de ces sous-bases regroupant tous les documents d'une même catégorie. Les expérimentations réalisées sur un jeu de dépêches de référence (Reuters-21578) montrent un certain retrait du catégoriseur proposé. Le classifieur obtenu à partir des règles d'association nécessite donc d'être encore amélioré afin de pouvoir rivaliser avec des méthodes comme les SVM. Tout d'abord, l'indication donnée par la mesure de confiance s'avère insuffisante. Il serait intéressant d'utiliser alors d'autres mesures de sélection comme notamment l'intensité d'implication pour sélectionner les règles les plus adaptées.

• [Jaillet *et al.*, 2003] Simon Jaillet, Maguelonne Teisseire, Jacques Chauché, et Violaine Prince. Classification automatique de documents : Le coefficient des deux écarts. Actes du congrés *INFORSID'2003*, pages 87–102, 2003.

• [Liu *et al.*, 1998] Bing Liu, Wynne Hsu, et Yiming Ma. Integrating classification and association rule mining. In *Knowledge Discovery and Data Mining*, pages 80–86, 1998.