Le Text Mining et XML

Nicole Munyana
30024815

RÉSUMÉ

L'extraction des termes complexes est une étape de prétraitement importante pour des opérations informatiques ou d'informatique linguistique complexes comme : la terminologie, le résumé automatique, mais aussi le text-mining et la classification textuelle. Nous présentons dans ce mémoire un filtre linguistique pour l'extraction des termes complexes. Ce filtre linguistique est fondé sur un modèle catégoriel : la Grammaire Catégoriel Combinatoire Applicative.

C'est grâce à ce modèle catégoriel que nous avons pu identifier les termes complexes à partir d'une liste des termes candidats. Une analyse syntaxique des formes phénotypiques qui est obtenue par un calcul sur les types syntaxiques, nous a permis de vérifier si un terme candidat est du groupe nominal. Ce sont les termes candidats ayant comme catégorie le groupe nominal qui sont préservés par notre filtre linguistique. Les termes candidats n'ayant pas comme catégorie le groupe nominal sont rejetés. L'expression applicative obtenue après l'analyse syntaxique a donné après réduction des combinateurs la forme normale du terme complexe. Les résultats du traitement des termes complexes sont stockés dans une base de données XML. XML offre des possibilités intéressantes pour des manipulations ultérieures de ces résultats par des requêtes XQuery.

Notre filtre linguistique est différent de la plupart des autres filtres linguistiques d'identification des termes complexes, parce qu'il tend à être multilingue. Avec la croissance du Web et des bases de données textuelles multilingues, cet aspect est significatif. Tout ce que nous avons besoin pour adapter l'approche à une nouvelle langue est un dictionnaire des types catégoriels avec les entrées lexicologiques de cette langue. L'approche théorique a été implémentée en C++ pour un corpus important du français.