Un modèle d'espace vectoriel de concepts pour noyaux sémantiques

Sujeevan Aseervatham*

*LIPN - UMR 7030
CNRS - Université Paris 13
99, Av. J.B. Clément
F-93430 Villetaneuse, France
Sujeevan.Aseervatham@lipn.univ-paris13.fr

Résumé. Les noyaux ont été largement utilisés pour le traitement de données textuelles comme mesure de similarité pour des algorithmes tels que les Séparateurs à Vaste Marge (SVM). Le modèle de l'espace vectoriel (VSM) a été amplement utilisé pour la représentation spatiale des documents. Cependant, le VSM est une représentation purement statistique. Dans ce papier, nous présentons un modèle d'espace vectoriel de concepts (CVSM) qui se base sur des connaissances linguistiques a priori pour capturer le sens des documents. Nous proposons aussi un noyau linéaire et un noyau latent pour cet espace. Le noyau linéaire exploite les concepts linguistiques pour l'extraction du sens alors que le noyau latent combine les concepts statistiques et linguistiques. En effet, le noyau latent utilise des concepts latents extraits par l'Analyse Sémantique Latente (LSA) dans le CVSM. Les noyaux sont évalués sur une tâche de catégorisation de texte dans le domaine biomédical. Le corpus Ohsumed, bien connu pour sa difficulté de catégorisation, a été utilisé. Les résultats ont montré que les performances de catégorisation sont améliorées dans le CSVM.

1 Introduction

Les mesures de similarité sont des éléments clés dans les algorithmes de traitement automatique des langues. Elles sont utilisées pour orienter le processus d'extraction de connaissance. Ainsi, elles sont les principales responsables des performances d'un algorithme. Si une mesure de similarité pertinente améliorera les performances, une mauvaise mesure risque de mener à des résultats incohérents. La définition d'une bonne mesure n'est pas un processus aisé. En effet, la mesure doit donner une bonne indication sur le degré de similarité entre deux documents. La notion de sémantique n'est pas clairement définie. Bien que nous essayons d'imiter la perception humaine, l'information sémantique peut prendre différente forme selon l'approche adoptée. Il existe deux grandes approches : l'une basée sur l'information statistique tel que la fréquence de co-occurrence des termes et l'autre basée sur des sources de connaissances externes telles que les ontologies.

Dans la communauté de l'apprentissage, les noyaux (Shawe-Taylor et Cristianini, 2004) sont utilisés depuis une décennie comme fonctions de similarité basées sur le cosinus formé