$ExpLSA$: utilisation d'informations syntaxico-sémantiques associées à LSA pour améliorer les méthodes de classification conceptuelle

Nicolas Béchet, Mathieu Roche, Jacques Chauché

Équipe TAL, LIRMM - UMR 5506, CNRS
Université Montpellier 2, 34392 Montpellier Cedex 5 - France
{nicolas.bechet,mroche,chauche}@lirmm.fr

Résumé. L'analyse sémantique latente (LSA - Latent Semantic Analysis) est aujourd'hui utilisée dans de nombreux domaines comme la modélisation cognitive, les applications éducatives mais aussi pour la classification. L'approche présentée dans cet article consiste à ajouter des informations grammaticales à LSA. Différentes méthodes pour exploiter ces informations grammaticales sont étudiées dans le cadre d'une tâche de classification conceptuelle.

1 Introduction

Le domaine de la classification de données textuelles se décline en de nombreux axes parmi lesquels la classification conceptuelle. Cette dernière consiste à regrouper des termes dans des concepts définis par un expert. Citons par exemple les termes *pot d'échappement*, *pare-brise* et *essuie glace* qui peuvent être classés dans le concept *automobile*. Afin d'établir une telle classification sémantique, la proximité de chacun des termes issus des textes doit être mesurée. Ces termes sont ensuite classés en fonction de leurs proximités sémantiques par un algorithme de fouille de données tels que les *Kppv (K plus proches voisins)* ou bien les *K moyennes* (Cornuéjols et Miclet (2002)).

Nous nous focalisons dans cet article sur la première étape de la réalisation d'une classification conceptuelle : l'étude de la proximité des termes. Afin de calculer une telle proximité, nous nous appuyons sur une méthode appelée Latent Semantic Analysis (LSA) développée par Landauer et Dumais (1997)[1]. La méthode LSA est uniquement fondée sur une approche statistique appliquée à des corpus de grande dimension consistant à regrouper les termes (classification conceptuelle) ou les contextes (classification de textes). Une fois l'analyse sémantique latente appliquée à un corpus, un espace sémantique associant chaque mot à un vecteur est retourné. La proximité de deux mots peut alors être obtenue par un calcul de similarité comme le cosinus entre deux vecteurs. L'objectif de nos travaux est d'améliorer les performances de LSA par une approche nommée *ExpLSA* (***Exp***ansion des contextes avec ***LSA***).

L'approche *ExpLSA* consiste à enrichir le corpus qui constituera l'entrée d'une analyse sémantique latente *classique*. Cet enrichissement utilise les informations sémantiques obtenues

[1] voir aussi, http ://www.msci.memphis.edu/∼wiemerhp/trg/lsa-followup.html