Adéquation des modèles de représentation aux méthodes de catégorisation

Simon Jaillet*, Maguelonne Teisseire*
Gérard Dray**

*LIRMM-CNRS - ISIM-Université Montpellier 2
161 rue Ada, 34392 Montpellier Cedex 5 France
{jaillet, teisseire}@lirmm.fr
**LGI2P, EMA Site EERIE,
Parc Scientifique G. Besse, 30319 Nimes France
gerard.dray@ema.fr

Résumé. Cet article s'interesse à la problématique de la catégorisation de documents et plus particulièrement à l'impact de la méthode de représentation des documents dans le processus de catégorisation. A partir de différents jeux de documents représentés dans un espace vectoriel tout d'abord basé sur les concepts puis basé sur une approche de type $TF\text{-}IDF$, nous évaluons les méthodes de catégorisation SVM et Rocchio. Nous comparons ensuite les deux méthodes précédentes avec une méthode de clustering flou. Nous dressons ensuite le bilan des différentes représentations des textes en terme de qualité des résultats de classification.

1 Introduction

Les documents numériques disponibles sont en nombre perpétuellement croissant. L'intérêt de disposer de méthodes, de techniques efficaces de classification n'est plus à démontrer et de nombreux travaux de recherche [Sebastiani, 2002, Yang et Liu, 1999] se focalisent sur cet aspect. Les résultats obtenus sont utiles aussi bien pour la recherche d'information que pour l'extraction de connaissance. L'objectif est de classer de façon automatique les documents dans des catégories qui ont été définies soit préalablement par un expert, il s'agit alors de classification supervisée ou catégorisation, soit de façon automatique, il s'agit alors de classification non supervisée ou encore clustering. De façon très globale, le processus de catégorisation de document peut être décomposé selon : (1) une étape de formalisation textuelle des documents, (2) une étape d'apprentissage.

Dans cet article, nous nous intéresserons plus particulièrement à l'impact de la représentation des documents textuels (et des catégories) sur les différents algorithmes d'apprentissage (supervisés ou non). Il existe de nombreuses représentations textuelles cependant la plus utilisée est la représentation statistique de type $TF\text{-}IDF$ où chaque dimension de l'espace vectoriel correspond à un élement textuel, nommé terme d'indexation. Dans [Jaillet *et al.*, 2003], les documents sont représentés non plus en fonction des mots qu'ils contiennent mais en fonction d'une projection de ces derniers sur un ensemble fini de concepts. L'objectif est d'intégrer plus de sémantique dans la modélisation des documents. Mais enrichir les données manipulées ne permet pas