SOM pour la Classification Automatique Non supervisée de Documents Textuels basés sur Wordnet

Abdelmalek Amine*, Zakaria Elberrichi*, Michel Simonet **, Mimoun Malki*

* Laboratoire EEDIS, Département d'informatique, UDL, Sidi bel Abbes – Algérie
amine_abd1@univ-sba.dz, elberrichi@univ-sba.dz, malki_m@univ-sba.dz
** Laboratoire TIMC-IMAG, IN3S, Université Joseph Fourier, Grenoble - France
michel.simonet@imag.fr

Résumé. Dans cet article, nous proposons la méthode des SOM (cartes auto-organisatrices de Kohonen) pour la classification non supervisée de documents textuels basés sur les n-grammes. La même méthode basée sur les synsets de WordNet comme termes pour la représentation des documents est étudiée par la suite. Ces combinaisons sont évaluées et comparées.

1 Introduction

Mettre en œuvre l'une des méthodes de classification non supervisée consiste en premier lieu à choisir une manière de représenter les documents (Sebastiani, 2002) ; dans un second temps il faut choisir une mesure de similarité, et en dernier lieu choisir un algorithme de classification que l'on va mettre au point à partir des descripteurs et de la métrique choisis. Tout document d_j sera transformé en un vecteur de poids w_{kj} des termes t_k. La majorité des méthodes, pour calculer le poids w_{kj}, sont axées sur une représentation vectorielle des textes de type *TF-IDF* (Sebastiani, 2002), qui attribue un poids d'autant plus fort que le terme apparaît souvent dans le document et rarement dans le corpus complet. Il existe différentes approches pour la représentation des documents. Typiquement, la similarité entre documents est estimée par une fonction calculant la distance entre les vecteurs de ces documents. Plusieurs mesures de similarité ont été proposées (Jones & Furnas, 1987). Parmi ces mesures on peut citer la distance du cosinus. L'algorithme SOM (Kohonen & al, 2000) a été depuis longtemps proposé et appliqué dans le domaine de la classification des documents textuels. Cependant, les combinaisons entre SOM et représentation conceptuelle de textes d'une part, SOM et représentation basée sur les n-grammes d'autre part n'ont pas été beaucoup étudiées.

2 Expérimentations, résultats et évaluation

Les données utilisées dans nos expérimentations sont issues des textes du corpus Reuters21578. Dans l'approche basée sur les n-grammes, on compte les fréquences des n-grammes trouvés. Dans l'approche conceptuelle, on remplace les termes par les concepts qui leur sont associés dans l'ontologie de références lexicales Wordnet (Miller, 1990). Cette représentation nécessitera deux étapes : la première est le « mapping » des termes dans des concepts et le choix de la stratégie de « merging », la deuxième est l'application d'une stratégie de désambiguïsation. On choisit la stratégie « Concept seulement », où il s'agit de