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

remplacer le vecteur des termes par le vecteur des concepts en excluant tous les termes de la nouvelle représentation, y compris les termes qui n'apparaissent pas dans Wordnet. Pour la désambiguïsation nous utilisons la stratégie du « Premier concept » et la fonction *TFIDF* pour le calcul des poids de chaque terme pour les deux approches. Nous avons utilisé une carte de Kohonen 7x7. Pour chaque approche, quatre mesures de similarité ont été testées: les distances du cosinus, euclidienne, euclidienne au carré, et la distance de Manhattan. Nous avons calculé, pour chaque cas, le nombre de classes, le temps et le taux d'apprentissage. Nous avons pu observer que malgré les bons résultats obtenus par la méthode des n-grammes particulièrement pour n=3 et n=4, ceux obtenus par la méthode conceptuelle, avec la distance du cosinus, sont plus performant. Pour évaluer la qualité des classifications obtenues nous avons utilisé la f-mesure et l'entropie. La partition P considérée comme la plus pertinente et qui correspond le mieux à la solution externe attendue est celle qui maximise la F-mesure associée ou minimise l'entropie associée. La plus grande valeur de la f-mesure est **62,5** et la plus petite valeur de l'entropie est **37,5**. Ces deux valeurs correspondent à l'approche conceptuelle (Wordnet) avec la distance du cosinus, ce qui confirme les conclusions tirées.

3 Conclusion et perspectives

Dans cet article nous avons proposé deux nouvelles approches pour la classification non supervisée de textes, l'une basée sur l'utilisation des n-grammes et l'autre sur WordNet. Les résultats obtenus montrent que malgré les bons résultats obtenus par la méthode des n-grammes, le fait d'ajouter des connaissances lexicales dans la phase représentation permet de construire une classification de meilleure qualité. Nous projetons dans un premier temps d'utiliser d'autres stratégies de désambiguïsation et voir leur influence sur la classification, et dans un second temps utiliser d'autres approches conceptuelles de références syntaxiques pour la classification par la méthode SOM des textes multilingues.

Références

Jones, W. and Furnas, G. (1987). Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American Society for Information Science*, 38(6):420-442.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., Saarela, A. (2000). *Self organization of a massive document collection*. IEEE Transactions on Neural Networks. 11(3), May.

Miller, G.A. (1990). Wordnet: An on-line lexical database. InSpecial. *Issue of International Journal of Lexicography*, Vol 3, No.4, Chongqing, China.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1–47.

Summary

In this paper, we initially propose the method of the SOM(self-organizing maps of Kohonen) for unsupervised classification of textual documents based on the n-grams representation. The same method based on the synsets of WordNet as terms for the representation of documents is studied thereafter. These combinations are evaluated and compared.