Graphes des liens et anti-liens statistiquement valides entre les mots d'un corpus textuel : test de randomisation TourneBool sur le corpus Reuters

Alain Lelu* **, Martine Cadot** ***

*Université de Franche-Comté
30, rue Mégevand
25030 Besancon Cedex
Alain.Lelu@univ-fcomte.fr
**LORIA, Bât. C
Campus scientifique, BP 239
54506 Vandoeuvre lès Nancy Cedex
Alain.Lelu@loria.fr
** Université Henri Poincaré – Nancy1
Département informatique, BP 239
54506 Vandoeuvre lès Nancy Cedex
Martine.Cadot@loria.fr
http://www.loria.fr/~cadot/

Résumé. La définition du voisinage est un élément central en fouille de données, et de nombreuses définitions ont été avancées. Nous en proposons ici une version statistique issue de notre test de randomisation TourneBool, qui permet, à partir d'un tableau de relations binaires objets décrits / descripteurs, d'établir quelles relations entre descripteurs sont dues au hasard, et lesquelles ne le sont pas, sans faire d'hypothèse sur les lois de répartitions sous-jacentes, c'est-à-dire en tenant compte de lois de tous types sans avoir besoin de les spécifier. Ce test est basé sur la génération et l'exploitation d'un ensemble de matrices randomisées ayant les mêmes sommes marginales en lignes et colonnes que la matrice d'origine. Après une première application encourageante à un corpus textuel réduit, nous avons opéré le passage à l'échelle adéquat pour traiter des corpus textuels de taille réelle, comme celui des dépêches Reuters. Nous caractérisons le graphe des mots de ce corpus au moyen d'indicateurs classiques comme le coefficient de clustering, la distribution des degrés et de la taille des « communautés », etc. Une autre caractéristique de TourneBool est qu'il permet aussi de dégager les "anti-liens" entre mots, à savoir les mots qui « s'évitent » plus qu'attendu du fait du hasard. Le graphe des liens et celui des anti-liens seront caractérisés de la même façon.