Etude de textes par leur image

Hubert Marteau*, Alexandre Lefevre*, Nicole Vincent**

*Laboratoire d'Informatique, 64 av Jean Portalis, 37200 Tours
hubert.marteau@etu.univ-tours.fr
http://www.rfai.li.univ-tours.fr/RFAI/default.htm
**Laboratoire CRIP5-SIP, Université de Paris 5, 45 rue des Saints Pères,
75270 Paris Cedex 06
nicole.vincent@math-info.univ-paris5.fr
http://www.math-info.univ-paris5.fr/sip-lab/

Résumé. Nous proposons une méthode automatique de comparaison de textes reposant sur une technique de transformation d'un texte en une image de taille donnée et l'analyse à l'aide des outils de la géométrie fractale. Nous présentons une application à l'étude d'un corpus de 90 textes longs.

1. Introduction

De par la taille du corpus textuel existant et de par sa perpétuelle croissance, il est de plus en plus nécessaire de recourir à des outils de classification et d'indexation de textes. Nous proposons un système de représentation des textes lié à un système d'indexation tous les deux indépendants de la taille du corpus.

On peut distinguer trois types d'indexation : l'indexation manuelle, l'indexation semi-automatique (supervisée) et l'indexation automatique (non supervisée) [Pouliquen, 2002]. L'indexation manuelle nécessite qu'une personne désigne les termes d'indexation. Il est évident que ce type d'indexation n'est pas imaginable pour un passage à l'échelle. L'indexation semi-automatique propose à un utilisateur les termes d'indexation possibles selon leur fréquence, l'utilisateur n'a plus qu'à les accepter ou non. Enfin l'indexation automatique ne demande aucune participation de l'utilisateur. C'est évidemment ce dernier type d'indexation qui doit être mis en œuvre pour le traitement de gros corpus.

Nous proposons néanmoins, pour certaines applications, de faire reposer cette indexation automatique sur une phase d'apprentissage. Un des éléments, essentiel et dual de l'indexation, est la recherche qui repose sur la comparaison entre les représentations choisies pour l'indexation.

Il existe de nombreuses méthodes de mesure dans l'espace des unités textuelles et des mots [Lelu, 2002], nous pouvons en particulier citer le cosinus de Slaton [Salton, 1989], la distance du khi-deux, et le cosinus dans l'espace distributionnel.

Les méthodes d'indexation de textes se ramènent à un condensé d'informations des textes originaux : vecteurs, statistiques, par exemple. Le problème de cette nouvelle représentation est qu'elle n'admet aucune limite de taille prévisible. Les changements de représentations tels que les vecteurs, les statistiques, … permettent certes de condenser l'information, mais ne permettent pas d'obtenir une quantité fixe d'informations.

Afin de remédier à ce problème, nous proposons de transformer un texte en image, plus précisément d'associer une image de taille fixe à tout texte.