# Un automate pour évaluer la nature des textes

Hubert Marteau*, Nicole Vincent**

*Laboratoire d'Informatique, 64 av Jean Portalis, 37200 Tours
hubert.marteau@etu.univ-tours.fr
http://www.li.univ-tours.fr
**Laboratoire CRIP5-SIP, Université 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/crip5/

**Résumé.** On ne peut s'intéresser aux textes sans s'intéresser à leur nature. La nature des textes permet de distinguer les textes d'un point de vue primaire. Elle est utilisée pour identifier les textes artificiels, pour la reconnaissance de la langue, afin d'identifier les SPAMS … En ce sens, la méthode la plus connue reste encore la méthode de Zipf. Cet article propose une nouvelle méthode basée sur un automate. L'automate construit un signal pour chaque texte. L'automate est présenté en détail et des expérimentations montrent son utilité dans les domaines aussi divers que ceux cités précédemment.

## 1 Introduction

L'indexation de textes consiste à trouver une représentation vectorielle d'un texte. Ce vecteur contient les caractéristiques propres au texte et il est, la plupart du temps, utilisé pour permettre une recherche rapide de textes ou d'informations présentes dans les textes.

La représentation la plus commune des textes est le vecteur de fréquence Salton (1989). Le passage du corpus à une telle représentation crée donc une matrice à deux dimensions. Les colonnes représentent les documents. Les lignes représentent les différents mots du corpus. Chaque valeur de la matrice indique le nombre de fois où le mot apparaît dans le texte.

Cavnar et Trenkle (1994) utilisent les vecteurs de fréquence avec les n-grammes de caractères et non les mots. Labbé et Labbé (2001) étudient des fréquences normalisées. Ils espèrent ainsi ôter le biais amené par les différences de longueur des textes. Mothe et al. (2001) proposent, pour améliorer l'indexation, de définir une représentation en trois dimensions. La troisième dimension sert à représenter des informations de type structurel (balises). De Vel (2000) représente les textes essentiellement par leurs caractéristiques structurelles : nombre de total de mots, longueur moyenne des mots, nombre de phrases, …, fréquence d'apparition de mots outils.

L'indexation fait parfois intervenir des méthodes liées davantage au traitement de la langue naturelle Rajman et Besançon (2004). Ainsi, Da Sylva (2004) ne se contente pas uniquement des mots, mais ajoute à la représentation des termes (préfixes, suffixes, …), des paires de mots, …

Cependant d'autres choisissent une autre modélisation. SanJuan et Ibekwe-SanJuan (2002) représentent chaque texte sous forme d'un graphe. Ce graphe est construit à partir des