L'automate textuel pour la prise en compte de l'évolution du texte

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. Il n'est plus à rappeler que le corpus textuel, est tel qu'il est actuellement, intraitable à l'échelle et que sa croissance nous confirme l'obligation d'utiliser des outils automatiques de traitement. Cet article s'intéresse plus particulièrement à la caractérisation de textes et par là même à celle d'auteurs. A l'heure actuelle, toutes les méthodes existant travaillent sur un à-plat des textes traités. C'est-à-dire que ces méthodes travaillent sur le document fini, sans admettre qu'un cheminement existe entre le début du document et sa fin. Nous proposons une méthode tentant d'apporter cette notion d'évolution textuelle en traitant le texte par un automate. Cette méthode a pour but d'apporter des informations complémentaires quant au cheminement logique adopté lors de la création du texte. Nous présenterons les règles de l'automate et l'évaluation choisie. Puis nous présenterons des résultats validés par des experts, obtenus sur un corpus d'entretiens sociologiques.

1 Introduction

Le corpus textuel existant atteint une taille qui depuis longtemps le rend intraitable par l'homme de manière exhaustive. Sa perpétuelle croissance et le fait qu'aucune limite ne peut être envisagée à cette croissance confirment le besoin de traitement automatique de toute la quantité de données brutes présentes dans ce corpus. La première étape, comme pour tout problème de gestion de données, consiste à effectuer une classification des données.

La classification des données a pour but de créer un système dans lequel chaque donnée appartient à un ou plusieurs groupe(s) selon les informations que l'on souhaite traiter. (Mothe et al 2001) par exemple se limitent aux balises des textes traités, c'est-à-dire le titre, les auteurs, etc. … Il existe (Pouliquen 2002) trois types de classification : la classification manuelle, la classification semi-automatique et la classification automatique. La classification manuelle est le résultat d'un traitement par l'homme ; comme on l'a indiqué précédemment, elle est impossible à envisager. La classification 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. La classification automatique est réalisée de manière complètement automatique et a pour résultat, dans la plupart des cas, d'établir une distance entre les textes.

L'un des résultats les plus connus est sans doute celui de Salton (Salton 1971)(Salt 1989) et son vecteur de données (Singhal et Salton 1995).