Intégration de la structure dans un modèle probabiliste de document

Mathias Géry, Christine Largeron et Franck Thollard

Université Jean Monnet,
Laboratoire Hubert Curien, UMR CNRS 5516, St-Etienne
prenom.nom@univ-st-etienne.fr

Résumé. En fouille de textes comme en recherche d'information, différents modèles, de type probabiliste, vectoriel ou booléen, se sont révélés bien adaptés pour représenter des documents textuels mais, ces modèles présentent l'inconvénient de ne pas tenir compte de la structure du document. Or la plupart des informations disponibles aujourd'hui sur Internet ou dans des bases documentaires sont fortement structurées. Dans cet article[1], nous proposons d'étendre le modèle probabiliste de représentation des documents de façon à tenir compte du poids d'une certaine catégorie d'éléments structurels : les balises représentant la structure logique et la structure de mise en forme. Ce modèle a été évalué à l'aide de la collection de la campagne d'évaluation INEX 2006.

1 Introduction

En fouille de texte comme en recherche d'information (RI), plusieurs modèles sont utilisés pour représenter un document. Ces modèles, de type probabiliste, booléen ou vectoriel, se sont révélés bien adaptés pour représenter des documents textuels. Cependant, ils présentent l'inconvénient de ne pas tenir compte de la structure du document. Or, la plupart des informations disponibles aujourd'hui sur Internet ou dans des bases documentaires sont fortement structurées. C'est la raison pour laquelle des travaux récents, en RI comme en fouille de données se sont intéressés à la structure des documents. Ceci a notamment conduit à l'émergence de la recherche d'information XML orientée contenu dont l'objectif est justement d'exploiter l'information structurelle contenue dans les documents pour concevoir des systèmes de RI plus efficaces. La compétition INEX[2] (INitiative for Evaluation of XML Retrieval) produit d'ailleurs depuis 2002 de larges collections de documents utilisables pour l'évaluation de tels systèmes. L'exploitation de la structure a aussi été étudiée dans des tâches de classement, supervisé ou non, de documents . Dans ce contexte, plusieurs voies ont été envisagées, parmi lesquelles on citera l'extension des modèles usuels de représentation de documents textuels [Doucet et Ahonen-Myka (2002)] ou l'exploitation de la structure arborescente des documents XML [Yi et Sundaresan (2000); Marteau et al. (2005); Vercoustre et al. (2006)]. Enfin, dans le contexte de la détection d'information nouvelle (Novelty Detection), d'autres travaux ont

[1]Ce travail a été partiellement soutenu par l'action collaborative Web Intelligence de la région Rhône-Alpes
[2]http ://inex.is.informatik.uni-duisburg.de/2007/