Fouille de collections de documents en vue d'une caractérisation thématique de connaissances textuelles

Abdenour Mokrane, Gérard Dray, Pascal Poncelet

Groupe Connaissance et Systèmes Complexes
LGI2P – Site EERIE – EMA
Parc scientifique Georges Besse, 30035 Nîmes cedex 1 - France
Tél : +33 (0)4 66 38 70 94 Fax : +33 (0)4 66 38 70 74
{abdenour.mokrane, gerard.dray, pascal.poncelet}@ema.fr

Résumé. De nos jours, les entreprises, organismes ou individus se trouvent submergés par la quantité d'information et de documents disponibles. Les utilisateurs ne sont plus capables d'analyser ou d'appréhender ces informations dans leur globalité. Dans ce contexte, il devient indispensable de proposer de nouvelles méthodes pour extraire et caractériser de manière automatique les informations contenues dans les bases documentaires. Nous proposons dans cet article l'approche *IC-Doc* de caractérisation automatique et thématique du contenu de collections de documents textuels. *IC-Doc* est basée sur une méthode originale d'extraction et de classification de connaissances textuelles prenant en considération les co-occurrences contextuelles et le partage de contextes entre les différents termes représentatifs du contenu. *IC-Doc* permet ainsi une extraction automatique de *KDMs* (*Knowledge Dynamic Maps*) sur les contenus des bases documentaires. Ces *KDMs* permettent de guider et d'aider les utilisateurs dans leurs tâches de consultations documentaires. Ce papier présente également une expérimentation de notre approche sur des collections de documents textuels.

Mots-Clefs. Caractérisation thématique, Similarité textuelle, Partage de contextes, Knowledge Dynamic Map.

1 Introduction

La fouille de données textuelles vise essentiellement à résoudre les problèmes de surabondance d'informations et faciliter l'extraction des connaissances enfouies dans les documents disponibles sur les bases de données ou sur le Web. Chaque jour, en particulier en raison de l'essor des communications électroniques, le nombre de documents disponibles croît de manière exponentielle et l'utilisateur (entreprise, organisme ou individu) se trouve submergé par la quantité d'informations disponibles. Ces utilisateurs ne sont donc plus capables d'analyser ou d'appréhender ces informations dans leur globalité.

De nombreux travaux de recherche, notamment issus du Web Mining et du Text Mining, s'intéressent aux traitements de bases de documents textuels (Baldi et Di meglio 2004,