Classification automatique de documents bruités à faible contenu textuel

Sami Laroum*, Nicolas Béchet**, Hatem Hamza*** et Mathieu Roche**

* Biopolymères Intéractions Assemblages, INRA
BP 71627, 44316 Nantes - France
sami.laroum@nantes.inra.fr,
** LIRMM Université Montpellier II CNRS
161 Rue Ada, 34392 Montpellier -France
{nicolas.bechet, Mathieu.Roche}@lirmm.fr,
http://www.lirmm.fr/
***ITESOFT: Parc d'Andron
Le Séquoia 30470 Aimargues -France
Hatem.Hamza@itesoft.com
http://www.itesoft.fr/

Résumé. La classification de documents numériques est une tâche complexe dans un flux numérique de gestion électronique de documents. Cependant, la quantité des documents issus de la retro-conversion d'OCR (Reconnaissance Optique de Caractères) constitue une problématique qui ne facilite pas la tâche de classification. Après l'étude et l'évaluation des descripteurs les mieux adaptés aux documents issus d'OCR, nous proposons une nouvelle approche de représentation des données textuelles : l'approche HYBRED (**HYB**rid **RE**presentation of **D**ocuments). Elle permet de combiner l'utilisation de différents descripteurs d'un texte afin d'obtenir une représentation plus pertinente de celui-ci. Les expérimentations menées sur des données réelles ont montré l'intérêt de notre approche.

1 Introduction

Aujourd'hui, nous vivons dans un monde où l'information est disponible en grande quantité tout en étant de qualité très diverse. Internet s'enrichit continuellement de nouveaux contenus. Par exemple, les entreprises emmagasinent de plus en plus de données, le courriel devient un moyen de communication extrêmement populaire, des documents autrefois manuscrits sont aujourd'hui disponibles sous format numérique. Mais toute cette information complexe serait sans intérêt si notre capacité à y accéder efficacement n'augmentait pas elle aussi. Pour cela, nous avons besoin d'outils permettant de chercher, classer, conserver, mettre à jour et analyser les données accessibles. Il est ainsi nécessaire de proposer des systèmes afin d'accéder rapidement à l'information désirée, réduisant ainsi l'implication humaine.

Un des domaines qui tente d'apporter des améliorations et de réduire la tâche de l'humain est la classification automatique de documents. Celle-ci consiste à associer une catégorie à un