Classification de documents XML à partir d'une représentation linéaire des arbres de ces documents

Anne-Marie Vercoustre*, Mounir Fegas*
Yves Lechevallier*, Thierry Despeyroux*

*INRIA Rocquencourt
B.P. 105 78153 Le Chesnay Cedex France
Prénom.Nom@inria.fr,
http://www-rocq.inria.fr

Résumé. Cet article présente un nouveau modèle de représentation pour la classification de documents XML. Notre approche permet de prendre en compte soit la structure seule, soit la structure et le contenu de ces documents. L'idée est de représenter un document par l'ensemble des sous-chemins de l'arbre XML de longueur comprise entre n et m, deux valeurs fixées a priori. Ces chemins sont ensuite considérés comme de simples mots sur lesquels on peut appliquer des méthodes standards de classification, par exemple K-means. Nous évaluons notre méthode sur deux collections: la collection INEX et les rapports d'activité de l'INRIA. Nous utilisons un ensemble de mesures bien connues dans le domaine de la recherche d'information lorsque les classes sont connues a priori. Lorsqu'elles ne sont pas connues, nous proposons une analyse qualitative des résultats qui s'appuie sur les mots (chemins) les plus caractéristiques des classes générées.

1 Introduction

XML est devenu un standard pour la représentation et l'échange de données. Le nombre de documents XML échangés augmente de plus en plus, et la quantité d'information accessible aujourd'hui est telle que les outils, même sophistiqués, utilisés pour rechercher l'information dans les documents ne suffisent plus. D'autres outils permettant de synthétiser ou classer de larges collections de documents sont devenus indispensables.

Dans ce contexte, de nombreux travaux proposent des méthodes de classification, supervisées ou non, pour organiser ou analyser de larges collections de documents XML. (Denoyer et al. (2003)) combinent plusieurs fonctions d'affectation (classifiers) pour classer des documents XML multimédia, (Despeyroux et al. (2005)) identifient, pour une collection homogène donnée, les types d'éléments XML les plus pertinents pour un objectif de classification. La similarité entre documents peut être définie en étendant le modèle vectoriel pour tenir compte de la structure (Doucet et Ahonen-Myka (2002), Yi et Sundaresan (2000)), ou seulement à partir de la structure d'arbre des documents, selon l'objectif visé ou l'hétérogénéité de la collection. Ainsi, la similarité structurelle peut être basée sur la distance entre arbres (Francesca et al. (2003), Nierman et Jagadish (2002), Dalamagas et al. (2004)), ou sur la détection de