# Expériences de classification d'une collection de documents XML de structure homogène

Thierry Despeyroux*, Yves Lechevallier*
Brigitte Trousse**, Anne-Marie Vercoustre*

*Inria - Rocquencourt
B.P. 105 - 78153 Le Chesnay Cedex, France
**Inria - Sophia Antipolis
B.P. 93 - 06902 Sophia Antipolis, France
email : Prénom.Nom@inria.fr
http ://www-rocq.inria.fr/axis/

**Résumé.** Cet article présente différentes expériences de classification de documents XML de structure homogène, en vue d'expliquer et de valider une présentation organisationnelle pré-existante. Le problème concerne le choix des éléments et mots utilisés pour la classification et son impact sur la typologie induite. Pour cela nous combinons une sélection structurelle basée sur la nature des éléments XML et une sélection linguistique basée sur un typage syntaxique des mots. Nous illustrons ces principes sur la collection des rapports d'activité 2003 des équipes de recherche de l'Inria en cherchant des groupements d'équipes (Thèmes) à partir du contenu de différentes parties de ces rapports. Nous comparons nos premiers résultats avec les thèmes de recherche officiels de l'Inria.

## 1 Introduction

Les documents XML sont maintenant incontournables et la classification de ces documents est un domaine de recherche très actif, en particulier pour définir des modèles de représentations de documents qui étendent les modèles traditionnels en tenant compte de la structure du texte (Yi and Dundaresan, 2000), (Denoyer and al.). Cela revient souvent à considérer que les même mots apparaissant dans des éléments XML différents sont en fait différents. Ces approches sont génériques, elles peuvent s'appliquer quelque soit la DTD, alors que notre approche suppose une connaissance d'une sémantique implicite des éléments pour les sélectionner.

Certaines méthodes de classification réduisent les documents XML à leur partie purement textuelle, sans prendre avantage de la structure qui pourtant véhicule une information riche. Nous nous intéressons à l'impact du choix des parties de documents sélectionnées sur le résultat de la classification, l'idée étant que ces différentes parties participent à différentes vues pouvant mener à des classifications différentes. Nous pratiquons successivement deux niveaux de sélection : une sélection utilisant la structure du document, puis une sélection linguistique au niveau du texte précédemment sélectionné. Nous utilisons ensuite un algorithme de classification qui va construire une partition des documents, affecter les documents à des classes et exhiber la liste des mots qui ont permis la classification.