# F-CheX : Une approche de fouille dans les documents XML

Amina MADANI*, Omar BOUSSAID**
Hafida ABED*

*Université Saad Dahlab de Blida (Algérie)
a_madani@univ-blida.dz, hafidabouarfa@hotmail.com

**Université Lumière Lyon2 (France)
Omar.Boussaid@univ-lyon2.fr
http://eric.univ-lyon2.fr/~boussaid/

**Résumé.** Nous présentons dans cet article une approche de fouille dans les documents XML qui prend en compte la structure et le contenu. Notre approche consiste à effectuer un *clustering* sur les documents XML. Ces derniers sont représentés par des ensembles de chemins conservant la structure arborescente des éléments. Les ensembles de chemins sont mappés dans une matrice sur laquelle une méthode de *clustering* est appliquée. L'approche proposée utilise un thésaurus créé au préalable pour gérer l'aspect sémantique des mots. Une évaluation de notre approche est effectuée à travers une étude expérimentale sur deux collections de documents XML.

## 1 Introduction

Le développement du document électronique et du web ont permis l'émergence de formats semi-structurés permettant la représentation et le stockage de documents textuels ou multimédias. Différents formats comme le XML sont aujourd'hui très populaires et sont en train de s'imposer. Ils permettent de représenter l'information sous une forme enrichie et adaptée à des besoins spécifiques. Ces types de formats permettent de représenter conjointement les données et une information sur leur structure.

En outre, les techniques d'analyse traditionnelles ne permettent pas une exploration adaptée à ce genre de documents. En effet, ces derniers contiennent des données semi-structurées et des métadonnées qui sont des informations sur celles-ci. Les techniques de fouille de données (*data mining*), plus particulièrement le *text mining*, ne tiennent pas compte des aspects spécifiques des documents XML.

Différents travaux sur le *XML mining* sont alors proposés. D'une manière générale, la fouille des documents XML s'appuie sur les techniques classiques de fouille et notamment le classement (*classification*) et la classification (*clustering*). La fouille des documents XML s'avère être un volet de recherche récent et assez peu exploré.

Dans cet article, nous présentons une approche de fouille des documents XML que nous proposons. Notre approche prend en considération le contenu textuel, qui représente le