Vers une méthode automatique de construction de hiérarchies contextuelles

Dino Ienco[*****], Yoann Pitarch[**],
Pascal Poncelet[***],
Maguelonne Teisseire[*****]

[*]Irstea, UMR TETIS, 500 rue Jean-Francois Breton, F-34093 Montpellier, France
{Dino.Ienco,Maguelonne.Teisseire}@teledetection.fr,
[**] Département Informatique, Université d'Aalborg, Dk-9000 Aalborg, Danemark
ypitarch@cs.aau.dk
[***]LIRMM, 161 rue Ada, F-34090 Montpellier, France
Pascal.Poncelet@lirmm.fr

Résumé. Dans de nombreux domaines (*e.g.*, fouille de données, entrepôts de données), l'existence de hiérarchies sur certains attributs peut être extrêmement utile dans le processus analytique. Toutefois, cette connaissance n'est pas toujours disponible ou adaptée. Il est alors nécessaire de disposer d'un processus de découverte automatique pour palier ce problème. Dans cet article, nous combinons et adaptons des techniques issues de la théorie de l'information et du clustering pour proposer une technique *orientée données* de construction automatique de taxonomies. Les deux principaux avantages d'une telle approche sont son caractère totalement non-supervisé et l'absence de paramètre utilisateur à spécifier. Afin de valider notre approche, nous l'avons appliquée sur des données réelles et avons conduit plusieurs types d'expérimentation. D'abord, les hiérarchies obtenues ont été expertisées pour en examiner le pouvoir informatif. Ensuite, nous avons évalué l'apport de ces taxonomies comme support à des tâches de fouille de données nécessitant une définition hiérarchique des valeurs d'attributs : l'extraction de séquences fréquentes multidimensionnelles et multi-niveaux ainsi que la construction de résumés de tables relationnelles. Les résultats obtenus permettent de conclure quant à l'intérêt de notre approche.

1 Introduction

Les taxonomies fournissent une organisation hiérarchique des données couramment exploitée dans de nombreux domaines tels que la biologie (Ashburner et al. (2000)), le e-commerce (Kohavi et Provost (2001)), le web sémantique (T. Berners-Lee et Lassila (2001)). Elles permettent de réaliser différents types d'analyse de données telles que l'anonymisation (Samarati (2001)), l'exploration d'entrepôts de données (Pitarch et al. (2010)) ou le résumé de données (Candan et al. (2010)). Par exemple, produire un résumé de données peut être particulièrement utile pour un décideur qui souhaiterait obtenir un rapide aperçu des ventes à l'échelle nationale plutôt que de considérer les ventes individuelles de chaque magasin. Traditionnellement,