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.

Summary

In many domains, a hierarchical organization of attribute values can help the data analysis process. Nevertheless, such hierarchical knowledge does not always available or even may be inadequate or useless when exists. Starting from this consideration, in this paper we tackle the problem of the automatic definition of data-driven taxonomies. To do this we combine techniques coming from information theory and clustering to obtain a structured representation of the attribute values: the Contextual Attribute-Value Taxonomy (CAVT). The two main advantages of our method are to be fully unsupervised and parameter-free. We experiments the

benefit of use CAVTs in the two following tasks: (i) multidimensional sequential pattern mining problem in which hierarchies are needed, (ii) table summarization problem, in which the hierarchies are used to aggregate the data. To validate our approach we use real world datasets in which we obtain appreciable results regarding both quantitative and qualitative evaluation.