Sujet de master recherche « Architectures logicielles distribuées »
2005–2006

Évaluation de la qualité des collections de résumés : application à la maintenance et à l'alignement de résumés

Encadrant principal : Guillaume Raschia
courriel : `Guillaume.Raschia@univ-nantes.fr`
tél. : 02 40 68 32 57

Co-encadrant(s) : Noureddine Mouaddib

Objectif du stage

L'équipe Atlas-GRIM a développé un modèle pour la réduction sémantique de données [1]. Les résumés ainsi définis représentent un ensemble d'enregistrements d'une table relationnelle et les décrivent de façon intentionnelle en utilisant des connaissances de domaine matérialisées par des caractérisations symboliques floues. Une des motivations essentielles de l'approche est de construire une représentation réduite des données tabulaires à l'aide d'une ou plusieurs collections de résumés [2]. Le problème qui suit immédiatement réside dans le *meilleur* choix de la collection de résumés pour cette réduction. De façon analogue, ce problème est très bien connu en analyse de données dans le cadre d'algorithmes de classification et a donné lieu à une littérature abondante sur la notion de similarité [3]. Il est donc pour nous indispensable de définir et de mettre en œuvre une mesure d'évaluation de la qualité des collections de résumés capable de discriminer différentes configurations pour ne retenir que la meilleure.

Nous avons déjà travaillé sur le sujet et proposé un score de qualité pour des collections de résumés, fondé sur les notions duales de contraste et de typicité. L'analyse critique de cette mesure constitue le point de départ du stage et doit faire émerger ses limites au regard des propriétés attendues des résumés.

Bien entendu, cette mesure est au cœur de nombreux traitements parmi lesquels :

- la maintenance des résumés en fonction des mises à jour sur les données, de sorte à garantir la correction des résumés à tout instant ;
- l'alignement de résumés dans le cadre de la fusion de sources de données distribuées possédant chacune une représentation réduite sur un schéma de données commun.

Ces deux cas de figure doivent permettre d'éprouver la mesure retenue pour évaluer la qualité des collections de résumés.

Travail à réaliser

Á partir de la sémantique du modèle de résumés de bases de données développé dans l'équipe Atlas-GRIM, étudier en détail une ou plusieurs propositions pour une mesure d'évaluation de la qualité des collections de résumés. Appliquer la méthode d'une part sur un procédé de maintenance des résumés, et d'autre part sur un algorithme d'alignement de résumés utilisé en environnement multi-sources.

Références

[1] R. Saint-Paul. *Une architecture pour le résumé en ligne de données relationnelles et ses applications*. PhD thesis, Adv. N. Mouaddib and G. Raschia, University of Nantes, France, july 2005.

[2] R. Saint-Paul, G. Raschia, and N. Mouaddib. General purpose database summarization. In *Int. Conf. on Very Large Databases (VLDB 2005)*, pages 733–744, Trondheim, Norway, 2005. Morgan Kaufmann Publishers.

[3] G. Bisson. La similarité : une notion symbolique/numérique. *Apprentissage symbolique-numérique (tome 2)*, pages 169–201, 2000. Eds Moulet, Brito. Editions CEPADUES.