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

Analyse en ligne de résumés de bases de données

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

Co-encadrant(s) : Noureddine Mouaddib

Objectif du stage

Il s'agit d'étudier, de concevoir et de développer un moteur de manipulation d'objets complexes appelés résumés. Les résumés [1] sont des versions réduites de données tabulaires, dont le modèle tire profit de la théorie des sous-ensembles flous pour capturer l'incertitude, l'imprécision et la gradation dans des descriptions intentionnelles de collections de n-uplets.

Étant donnée une table relationnelle, nous avons mis au point un algorithme incrémental de calcul de résumés qui génère plusieurs réductions complètes de la table, avec des taux de compression variables [2]. Chaque forme réduite (i.e. une collection de résumés) de la table est liée aux autres par une relation de subsomption possédant un unique plus grand élément (un résumé capturant toute l'information de la table, équivalent à la racine de l'arbre construit à partir de l'ordre partiel sur les résumés). Les feuilles de l'arbre constituent les résumés de granularité la plus fine, i.e. couvrant peu de n-uplets. La navigation le long des branches de l'arbre permet de réaliser un compromis entre la précision de la version réduite et sa concision.

Une des pistes de travail poursuivie actuellement dans l'équipe Atlas-GRIM au travers de la thèse de doctorat de Lamiaa Naoum, consiste à proposer une algèbre de manipulation en ligne des résumés, à la manière des traitements réalisables sur les cubes de données OLAP [3, 4, 5]. Une bibliographie abondante sur le sujet est référencée ici [6]. Des résultats significatifs ont déjà été produits [7], et notamment la définition d'un modèle de données et d'un ensemble d'opérateurs constituant le cœur de cette algèbre. On y trouve une extension d'opérateurs relationnels (sélection, projection, jointure), des opérateurs structurels (permutation, arrangement), des opérateurs ensemblistes (union, intersection, produit cartésien, différence) et des opérateurs de granularité (zoom avant/arrière).

Travail à réaliser

La suite de ces travaux doit nous permettre de valider certaines hypothèses sur la sémantique des opérateurs (par exemple, établir la correction des transformations vis-à-vis des n-uplets) et d'étendre le pouvoir d'expression de l'algèbre en proposant des opérateurs inédits avec un souci constant de fournir à l'utilisateur final une somme d'outils effectifs pour l'analyse en ligne des résumés. Là encore, les propriétés de l'algèbre devront être préservées (fermeture, complétude, etc.).

L'étude détaillée de l'algèbre doit s'accompagner du développement d'un moteur d'analyse en ligne des résumés, dans lequel seront implémentés et validés expérimentalement les opérateurs précédemment étudiés.

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] Rakesh Agrawal, A. Gupta, and Sunita Sarawagi. Modeling multidimensional databases. In Alex Gray and Per-Åke Larson, editors, *Proc. 13th Int. Conf. Data Engineering, ICDE*, pages 232–243. IEEE Computer Society, 7–11 April 1997.

[4] Jim Gray, Surajit Chaudhuri, Adam Bosworth, Andrew Layman, Don Reichart, Murali Venkatrao, Frank Pellow, and Hamid Pirahesh. Data cube : A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *Data Mining and Knowledge Discovery*, 1(1) :29–53, 1997.

[5] Luca Cabibbo and Riccardo Torlone. Querying multidimensional databases. In *6th Int. Workshop on Database Programming Languages (DBPL)*, pages 319–335, 1997.

[6] Data warehousing and OLAP : A research-oriented bibliography. http ://www.daniel-lemire.com/OLAP/.

[7] L. Naoum, G. Raschia, and N. Mouaddib. Manipulating fuzzy summaries of databases : Unary operators and their properties. In *Joint EUSFLAT & LFA Conference (EUSFLAT-LFA 2005)*, Barcelona, Spain, 2005.