Mining Frequent Queries in Star Schemes

Tao-Yuan Jen*, Dominique Laurent*
Nicolas Spyratos**, Oumar Sy***
*LICP, Université de Cergy-Pontoise, 95302 Cergy-Pontoise Cedex, FRANCE
{tao-yuan.jen,dominique.laurent}@dept-info.u-cergy.fr
**LRI, Université Paris 11, 91405 Orsay Cedex, FRANCE
spyratos@lri.fr
***Université Gaston Berger, Saint-Louis, SENEGAL
oumar.sy@ugb.sn

Résumé. L'extraction de *toutes* les requêtes fréquentes dans une base de données relationnelle est un problème difficile, même si l'on ne considère que des requêtes conjonctives. Nous montrons que ce problème devient possible dans le cas suivant : le schéma de la base est un schéma en étoile, et les données satisfont un ensemble de dépendances fonctionnelles et de contraintes référentielles. De plus, les schémas en étoile sont appropriés pour les entrepôts de données et que les dépendances fonctionnelles et les contraintes référentielles sont les contraintes les plus usuelles dans les bases de données. En considérant le modèle des instances faibles, nous montrons que les requêtes fréquentes exprimées par sélection-projection peuvent être extraites par des algorithmes de type Apriori.

1 Introduction

The general problem of mining *all* frequent queries in a (relational) database, *i.e.,* all queries whose answer has a cardinality above a given threshold, is known to be intractable, even if we consider conjunctive queries only (Goethals 2004).

However, mining all frequent queries from a database allows for the production of relevant association rules that cannot be obtained by other approaches, even when dealing with multiple tables, such as in (Dehaspe and Raedt 1997; Diop *et al.* 2002; Faye *et al.* 1999; Han *et al.* 1996; Meo *et al.* 1997; Turmeaux *et al.* 2003). This is so because, in these approaches, association rules are mined in the *same* table. On the other hand, when mining all frequent queries, it is possible to obtain rules whose left and right hand sides are frequent queries mined in *different* tables. The following example, that serves as a running example throughout the paper, illustrates this point.

Example 1 *Let Δ be a database containing three tables, Cust, Prod and Sales, dealing with customers, products and sales transactions, respectively, and suppose that :*

- *the table Cust is defined over the attributes Cid, Cname and Caddr, standing respectively for the identifiers, the names and the addresses of customers,*
- *the table Prod is defined over the attributes Pid and Ptype, standing respectively for the identifiers and the types of products,*
- *the table Sales is defined over the attributes Cid, Pid and Qty where Qty stands for the quantity of a product bought by a customer.*