Ontology-Assisted Query Formulation in Multidimensional Association Rule Mining

[1] Chin-Ang Wu
Department of Information Management, I-Shou University, Kaohsiung County, Taiwan 840, R.O.C.
cwu@csu.edu.tw

Wen-Yang Lin,
Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan 811, R.O.C.
wylin@nuk.edu.tw

Chuan-Chun Wu
Department of Information Management, I-Shou University, Kaohsiung County, Taiwan 840, R.O.C.
miswucc@isu.edu.tw

Abstract

In the information era, the development of various electronic information resources have dramatically grown, mining useful information from large databases has become one of the most important issues in information research for users. Information technologies have provided many applicable solutions, yet there are still many problems that cause users to spend extra time to get real knowledge. In this paper, we show an ontology based system framework for multi-dimensional association rule mining that incorporates ontologies in order to help users develop correct queries, reduce the system resource consumption and improve the efficiency of the mining process.

1. Introduction

In the knowledge economics era, the efficient use of data to promote business competition and opportunities is an important movement. Information technologies have provided many applicable solutions, but many decision analyses are still greatly dependent on professionals' manual judgments. We expect that a good system should provide efficient, convenient and thorough mining mechanisms, as well as meet the users' professional demands. In other words, developing a system environment that provides users a way to repeatedly discover important rules or facts based on their professional judgment is absolutely an issue that research workers in the knowledge mining fields can not avoid. Currently, the research on the integration of data mining and data warehousing are mostly concentrated on data mining from data cube or multi-dimensional database. J. Han's research group pioneered this research subject. They combined OLAP and data mining to develop DBMiner, a system that provides minings of association rules, classification, prediction and clustering from data cube[3]. To provide OLAP with efficient data storage and data analysis, star schema, proposed by Kimball [5], is a prevalent data model being used in many data warehouse systems. But the hierarchical and other semantic relationships among attributes which may be helpful to the data mining systems are not able to be displayed in the star schema and hence leave data mining with limitations. Thus an ontology, which is used to describe and represent domain knowledge [4], can be applied to collect the semantic knowledge of data warehouse. The main purpose of this paper is to show an ontology based system framework for multidimensional association rule mining that incorporates ontologies in order to help users develop correct queries and reduce the system resource consumption and improve the efficiency of the mining process.

The rest of this paper is organized as follows. In section 2, we introduce multidimensional association rule mining and its query with constriants. The framework of ontology-based multidimensional association rule mining system is explained in section 3. In section 4, query formulation and checking with the help of ontologies are described. And finally this paper concluded in section 5.

2. Multidimensional association rule mining and its query with constraints

[1] Chin-Ang Wu is also a lecturer in Cheng Shiu University, Niaosong Township,Kaohsiung County 833.

0-7695-3032-X/07 $25.00 © 2007 IEEE
DOI 10.1109/GrC.2007.106

An association rule is an expression of the form $X \Rightarrow Y$, where X and Y are sets of items. Such a rule reveals that transactions in the database containing items in X tend to also contain items in Y. X is the *body* or the *antecedent* and Y is the *head* or the *consequent*. The probability, measured as the fraction of transactions that contain X also contain Y, is called the confidence of the rule. The support of the rule is the percentage of the transactions that contain all items in both X and Y.

A multidimensional association rule explores the association between data values from different dimensional attributes in data warehouses.

Definition 1. Consider a transaction table composed of k dimensions. Let x_{im} and y_{jn} be the values of dimensions X_i and Y_j, respectively. The form of a multi-dimensional association rule is:

$$X_1 = \text{“}x_{1_m}\text{”}, X_2 = \text{“}x_{2_m}\text{”}, \ldots, X_i = \text{“}x_{im}\text{”} \Rightarrow Y_1 = \text{“}y_{1_n}\text{”}, Y_2 = \text{“}y_{2_n}\text{”}, \ldots, Y_j = \text{“}y_{jn}\text{”}$$

Following the work in [9], we can divide multi-dimensional association rules into three different types: intra-dimensional association rule, inter-dimensional association rule, and hybrid association rule.

In multidimensional association rule mining, users need to decide what dimensions to use to group data and what interested data items to be projected for mining association rules. For example, if a user is interested in mining association rules among age_group, city and category from each customer's daily purchased, then "Customer_ID" and "Date" will be the grouping dimensions and "Age_group", "City" and "Category" will be the interested mining items. Table 1 is this example table for mining multidimensional association rules.

Definition 2. Suppose a star schema S containing a fact table F and m dimension tables $\{D_1, D_2, \ldots, D_m\}$. Let T be a jointed table from S composed of $a_1, a_2, \ldots, a_k$ attributes, such that $\forall a_i, a_j \in Attr(D_k)$, $1 \le i, j \le r$, $1 \le k \le m$. Here $Attr(D_k)$ denotes the attribute set of dimension table D_k. A constrained meta-pattern of multidimensional association rule mining from T is defined as follow,

$$MP: \langle t_G, t_M, ms, mc, [wc], [hc], [mr] \rangle,$$

where ms denotes the minimum support, mc the minimum confidence, t_G the group of transaction attributes, t_M the group of interested mining items, for $t_G, t_M \subseteq \{a_1, a_2, \ldots, a_k\}$. In addition to these four basic elements, the meta-pattern constraints include *wc* the where condition for data in the data warehouse, *hc* the having condition and *mr* the meta rule. The constraints of the meta-pattern are optional.

Table 1. An example table for multidimensional association rule mining

Grouping_ID		Interested Mining_Items		
Cust_ID	**Date**	**Age_group**	**City**	***Category**
C01	2007-02-01	21-30	Taipei	B,C,D,E
C02	2007-02-03	31-40	Tainan	A,D,E
C03	2007-02-10	31-40	Taichung	C,D
C01	2007-02-10	21-30	Taipei	A,E
C05	2007-02-12	41-50	Kaohsiung	A,B,C
C03	2007-02-15	31-40	Taichung	A,E
*A: Priter B: Laptop C: Desktop D: Memory E: Hard Disk				

A query of multidimensional association rule mining is defined as follows:

Mining multidimensional association rules
[**Metarule** <meta-rule>]
Grouping_ID <attribute list>
Mining_Item <attribute list>
From <dw_name>
[**Where** <wcondition>]
[**Having** <hcondition>]
Threshold <ms, mc>

If a user is interested in mining associations of HP printer with age_group, city or other products from each customer's daily transaction in the market of Japan with minimum support 30% and minimum confidence 50%, the example query is as follow.

Mining multidimensional association rules
Metarule Prodcut.Prod_Name="HP printer", $B_1,..,B_n \Rightarrow H_1,..,H_n$.
Grouping_ID Customer,date,
Mining_Item Age_group,City,Prod_Name
From Sale_Star
Where Country='Japan'
Threshold ms=30%, mc=50%

3. The framework of ontology-based multidimensional association rule mining system

It is well known that the data mining is a knowledge intensive process that requires domain-specific knowledge (ontology) [1][2]. A typical star schema shows only the structural relationships among dimensions and attributes but lack of semantic connections among them.

Therefore, we have proposed a data mining system framework that incorporates ontologies for mining association rules from data warehouse [6]. In this section we will describe and enhance this system framework to make query formulation and checking more specific. Figure 1 illustrates the system framework.

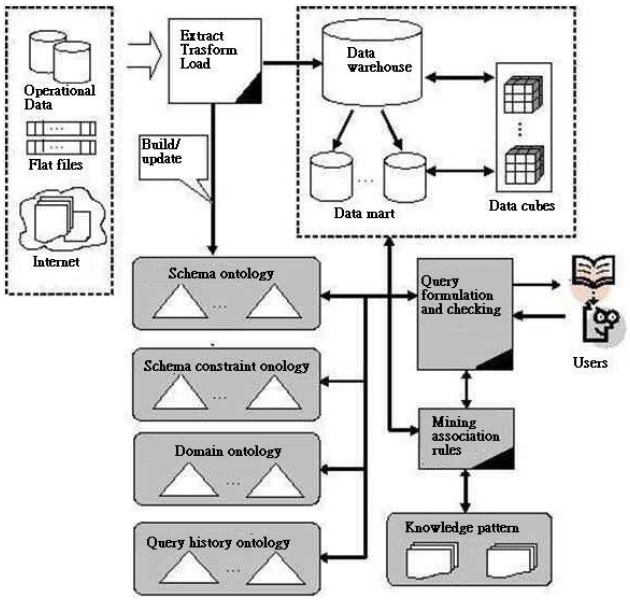

Figure 1. Ontology-based system framework for multidimensional association rule mining

The system framework includes a schema ontology, a schema constraint ontology, some domain ontologies and a user query history ontology. A user can form a mining query interactively in the system with the help of these ontologies. The query can then be processed by the association rule mining engine to generate knowledge patterns.

The schema ontology is used to describe metadata of the data warehouse, including schema structures, dimension hierarchies and the types of measures along dimensions. There exist some constraints of attributes in a data warehouse schema. In [7][8], the authors explored all possible mining spaces of grouping and mining attribute combinations. They then defined allowed range via the predicates functions, such as ignore, together, item-only, group-only, always-group, always-item, decide, follow and repel. These functions convey some constraints which are related to multidimensional association rule mining and these constraints are then recorded in the schema constraint ontology in our system to provide knowledge for users and the system as a reference to select grouping_ID and interested mining attributes. Figure 2 is an example of a schema constraint ontology.

A domain ontology is used to construct domain related knowledge of the mining subject. A domain ontology can contain relationships among classes such as classification and composition. Such knowledge can provide data mining with extra information beyond the data warehouse and hence reduce data mining searching boundaries and bring more innovative mining results.

The query history ontology is applied to collect history mining patterns and their corresponding results from users' mining activities. The elements of the meta-pattern defined in definition 2, the result rules and the result statistics will be included in the history ontology. Figure 3 is the graph of the query history ontology. The statistic data including number of frequent item set, average length of items, maximum length of items, minimum length of items, number of rules, average length of rules, maximum length of rules and minimum length of rules.

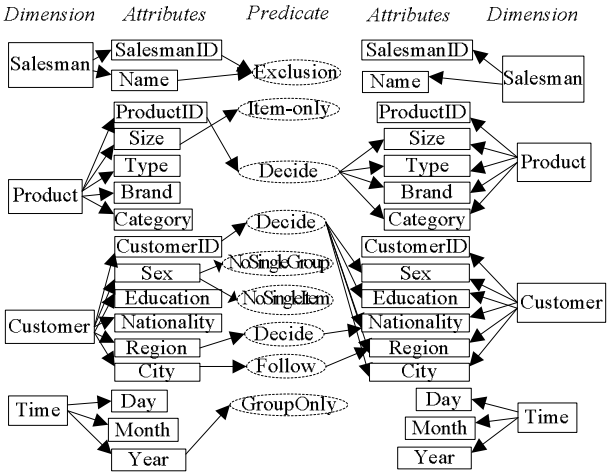

Figure 2. Schema constraint ontology

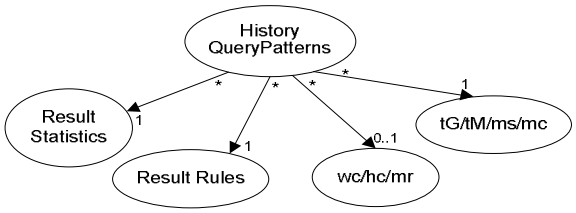

Figure 3. Query history ontology

4. Ontology assisted user query formulation

Without complete comprehension of the schema and domain related knowledge, end users may develop a query based on their experience or intuition. Therefore, users' formulation of queries can possibly fall into some improper pits. This may lead to incorrect and redundant mining data space or mining results and waste the efforts accordingly.

Through the assistance of ontologies the system will take more responsibilities in controlling the correctness and effectiveness of the query formulation. To get a correct query to feed into the mining process, the system will, after the user finish forming the query, first check the syntax and then perform the semantic checking if syntax check is all right. Figure 4 is the user query checking process.

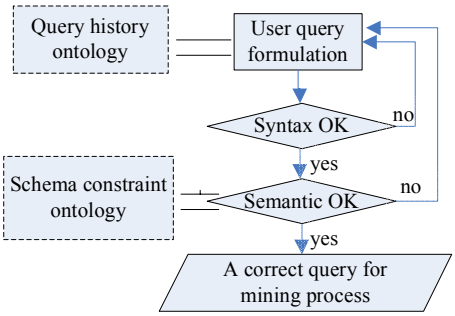

Figure 4. process of user query checking

While a user is formulating a query, the browse over query history ontology which provides users with abundant query format and result rules and statistics, can be performed to assist an inexperienced user with ideas of how to formulate a query. A user can also draw a query based on a history one or he or she can submit a query in order to compare the results with those have existed in the query history ontology. A query history ontology can therefore provide supplement information for users in order to generate useful queries that match better to the users need.

The system will check syntax of the query first and then progress toward semantic checking, which other systems disregard. While performing the semantic checking, the schema constraint ontology is a very important knowledge source for the system to refer. The schema constraint ontology provides restrictions on attribute combinations of group-by items and interested mining items. For example, the functional dependency relationship between Cust_ID and Sex will cause the mining query of t_G = {Cust_ID, Sex} and t_M = {Prod.Name} into a redundant form. In t_G, Cust_ID decides the value of Sex and therefore the groupedg tables of t_G = {Cust_ID, Sex} and t_G = {Cust_ID} will be the same. This kind of constraint is defined as decide(Cust_ID, Sex) in the schema constraint ontology. Note that a hierarchical relationship is a special case of functional dependency. The descendant's value determines its ancestor's value in the hierarchy.

There are varieties of other constraints that the system will utilize to check against the selected t_G and t_M. This can protect users from proposing improper mining queies.

5. Conclusion

In this paper, we have shown multidimensional association rule mining and its query with constraints and explained a system framework that incorporates with schema ontology, schema constraint ontology, domain ontology and query history ontology. The useful information left out by a data warehouse system can be accessible through the assistance of these ontologies. This system can help the user's query formulation and do query correctness checking in both syntax and semantics. Through the utilization of ontologies in the system framework, the correctness and efficiency of the mining process can be improved.

Acknowledgements

This work was supported by the National Science Council of R.O.C. under grant NSC 94-2213-E-390-006.

References

[1] J.M. Aronis, F.J. Provost, and B.G. Buchanan, "Exploiting background knowledge in automated discovery," in *Proc. 2nd Intl. Conf. Knowledge Discovery and Data Mining*, 1996.

[2] U. Fayyad, P.S. Gregory, and S. Padhraic, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, Vol. 39, No. 11, 1996, pp. 27–34.

[3] J. Han, "Mining knowledge at multiple concept levels," in *Proc. of ACM International Conference on Information and Knowledge Management*, 1995, pp. 19-24.

[4] Jeff Heflin (2004). OWL Web Ontology Language Use Cases and Requirements, http://www.w3.org/TR/webont-req/, Feb. 2004.

[5] R. Kimball, *The Data Warehouse Toolkit Practical For Building Dimensional Data Warehouses*, John Wiley & Sons, INC. 1996.

[6] Wen-Yang Lin, Chin-Ang Wu, Ming-Cheng Tseng, Chuan-Chun Wu (2006). "Ontology-Incorporated Mining of Association Rules in Data Warehouse", The 11th Conference on Artificial Intelligence and Applications, Kaohsiung, Taiwan, R.O.C.

[7] Chang-Shing Perng, Haixun Wang Ma, Joseph L. Hellerstein, "Farm: A Framework for Exploring Mining Spaces with Multiple Attributes," 2001 IEEE.

[8] Chang-Shing Perng, Haixun Wang Ma, Joseph L. Hellerstein, "User-directed exploration of Mining Space with Multiple Attributes," 2002 IEEE.

[9] Hua Zhu, "On-Line Analytical Mining of Association Rules," Master's Thesis, Simon Fraser University, U.S.A, Dec 1998.