Facilitating Active Multidimensional Association Mining with User Preference Ontology

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

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

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

Abstract

Multidimensional association mining from data warehouse has become a knowledge discovery paradigm because it provides more specific conditional settings for target mining data, thus can generate rules more close to users' needs. Yet data warehouse is subject to change by time or the modifications of business rules. Users might not know this change and reinitiate mining queries, which elicits the necessity of an active mining mechanism to bring new knowledge to users dynamically. In this paper, we propose an active multidimensional association mining system framework that incorporates with the user preference ontology that exploits frequent and representative queries. With the assistance of the user preference ontology and its association with user profile, the proposed system can facilitate active mining mechanism, allowing distribution of the renewal mining results to the specific users automatically.

1. Introduction

Data mining is to discover previous unknown knowledge from large amount of data, for the purpose of providing managers with useful information for decision making [2]. Data warehouse, proposed by W.H. Inmon [4], provides integrated data repository for data mining. It has solved the preprocessing problems of data mining; consequently, the users need not deal with issues such as data cleaning and integration. Nonetheless data warehouse grows through time and/or subject to change when the business rules change. New rules and new knowledge have to be extracted to reflect the most up-to-date situation. Periodic re-mining of rules hence is essential. Suppose that the data warehouse of a retailer is renewed once a month. Consider the following scenario: upon the time the data warehouse is loaded with new data of the month, the mining system automatically triggers the active mining process; the mining results are stored and the managers receive the newly discovered knowledge of their interest through the e-mail without delay. Thus the managers save the tedious work of re-mining process each month when new data comes. If each previous mining query is to be re-generated for the re-mining, it will be labor intensive and time consuming.

In this paper, we propose an active system framework of multidimensional association mining that utilizes user preference ontology in attempt to provide automatic triggering of a mining process without the involvement of a user's query formulation. The user preference ontology maintains the frequently used and representative queries in the mining history. The users' profiles such as e-mail, department, job title etc. are also combined in it. The system provides active triggering of the data re-mining based on the queries maintained in the user preference ontology and stores the resulting rules in a rule base. Specific re-mining results are dispatched automatically to specific users according to their preference.

The rest of this paper is organized as follows. Section 2 introduces the multidimensional association

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

mining. Section 3 presents the architecture of the user preference ontology. Section 4 describes the proposed active mining system framework. Section 5 summarizes related work and section 6 provides the conclusions.

2. Multidimensional association mining

An association rule has the form, A ⇒ B, where A and B are sets of items and A ∩ B = ∅. The rule implies that transactions in the data warehouse contain A tend to also contain B. For this rule to be interesting, A and B should exceed the user specified minimum support and minimum confidence. If only one attribute is involved in the rule, it is a single-dimensional association rule. If multiple attributes are involved in the rule, it is a multidimensional association rule. The following is an example multiple association rule involving two attributes, Education and Sub_category.

Education = "College", Sub_category = "Acer PC" ⇒ Sub_category = "HP printer".

The mining model of a multidimensional association mining can defined as follows:

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

where t_G, t_M, *wc*, *ms* and *mc* are components of a query, detailed below:

- t_G: the set of transaction ID (data granularity),
- t_M: the set of interested mining attributes,
- *wc*: the optional "*where*" condition(s),
- *ms*: the minimum support and
- *mc*: the minimum confidence.

In this model we see that it allows the users to set the data granularity which determines what the transaction ID(s) of the mining data is. Filtering condition is also allowed to let users set the specific range of the mining data he or she needs.

In general, the process of mining association rules starts with generating the frequent itemsets that satisfy the *ms* first and then derives the association rules based on the frequent itemsets that satisfy the *mc*.

3. The architecture of user preference ontology

In this study, we assume that for each mining process, the user's profile and the mining query are recorded in the mining log. The mining log is too tedious to be utilized in the active mining mechanism thus distillation for representative queries from it is necessary. We introduce the preference ontology to

The user preference ontology contains information in two aspects. First is the mining query and second is the user profile. The query distillation process involves the following steps as shown in Figure 1:

1. Find the query patterns of t_G and t_M that tend to be used together.
2. Determine the surrogate query patterns for query patterns of step 1.
3. Find the attributes (*wca*) in the "where" conditions that are likely to be used together with the surrogate query patterns of step 2.
4. Select queries with the "where" conditions from the query log based on the results of step 3.
5. Determine the surrogate "*where*" conditions for the "*where*" conditions of step 4.
6. Determine surrogate queries by combining the surrogate query patterns and the surrogate "where" conditions.
7. Calculate the minimum supports and minimum confidences of the surrogate queries.

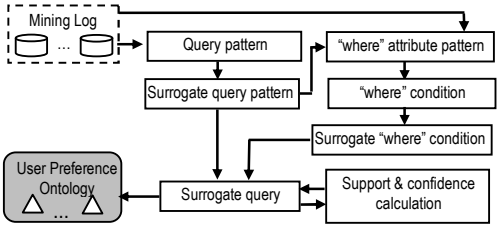

Figure 1. The distillation process

To obtain the query patterns that reveal the co-occurrence of t_G and t_M, we can, taking t_G and t_M as the interested mining attributes, apply any multi-dimensional association mining algorithm to the query log.

The step for determining a surrogate query pattern is better explained with examples. Suppose that the following query patterns conforming to the meta-rule format "$t_G \Rightarrow t_M$" are discovered:

QPattern1: t_G = {*CustID*}⇒ t_M = {*ProdName*}
QPattern2: t_G = {*CustID*}⇒ t_M = {*Education*}
QPattern3: t_G = {*CustID*}⇒ t_M = {*ProdName*, *Education*}

They indicate that if t_G = {*CustID*}, t_M tends to be {*ProdName*}, {*Education*}, {*ProdName*, *Education*}. To determine a surrogate query pattern, we have identified two principles to be followed. First, a surrogate query pattern represents patterns with the same t_G. Second, a surrogate query pattern covers the query patterns with the t_M that are its subsets. For the example above, QPattern1, QPattern2 and QPattern3

are closely related and QPattern3 is the surrogate query pattern because its t_M is a superset of that of QPattern1 and QPattern2. It is easy to verify that the rules discovered by a surrogate query patterns will include the rules that are to be discovered by all the query patterns that the surrogate query pattern represents. For example, the results of mining by QPattern3 will include that of QPattern1 and QPattern2.

In a multidimensional association mining, the "where" condition is allowed. We assume that the format of a "where" condition is

Where: (*conditional expression*) [*And/Or* (*conditional expression*)].

For example,

Where: *Country*="*Italy*" and *Sex*="*Female*"

This condition confines the interested mining data to be "*Italian female*". To determine whether a "where" condition is used often with t_G and t_M of a surrogate query pattern, we first extract the *wca* of a conditional expression and derive the patterns of the meta-rule format "t_G, $t_M \Rightarrow$ *wca*". The rules mined indicate that the users often specify the *attribute(s)* in "where" condition together with such t_G and t_M. For example,

SQPattern1: t_G = {*CustID*}, t_M = {*ProdName, Education*} $\Rightarrow$ *wca* = {*Country, Age*}
SQPattern2: t_G = {*CustID*}, t_M = {*ProdName, Education*} $\Rightarrow$ *wca* = {*Education*}

The surrogate "where" condition can be determined by analyzing the "where" conditions of the queries in the query log that have the t_G, t_M and *wca* in the SQPatterns. In order to represent other conditions, the query that covers the widest range is the surrogate query. For example, under certain t_G and t_M, suppose that the following "where" conditions are recorded in the query log for SQPattern1.

SQWhere1: Country = "Italy" and Age > "10~20"
SQWhere2: Country = "Italy" and Age >= "50~60"
SQWhere3: Country = "Italy" and Age >= "30~40"

Obviously, since the range exploited by SQWhere1 covers that of SQWhere2 and SQWhere3, SQWhere1 is the surrogate "where" condition for SQWhere2 and SQWhere3.

To calculate the *ms* and *mc* of the surrogate query, the average *ms* and *mc* of the original mining queries in the mining log with the same t_G, t_M and *wca* are obtained.

The user preference ontology is the concise version of the mining log that contains the frequently used queries with representative power.

The surrogate queries are structured and stored in the user preference ontology as shown in Figure 2. The attribute index provides fast access to t_G and t_M by associating the attribute to the corresponding t_G and t_M. The rules mined by a surrogate query should cover those mined from the queries it represents. A surrogate query is presented as a path in the user preference ontology. As such, the user preference ontology is structured with hierarchies in t_G — t_M — *wc*. It collects surrogate queries that are relatively used often before.

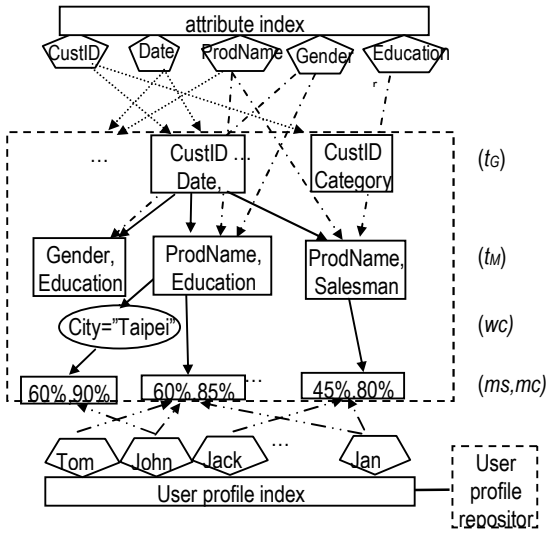

Figure 2. The architecture of user preference ontology

4. The Active Multidimensional Association Mining System Framework

The proposed system framework of the active multidimensional association mining that incorporates with the user preference ontology is shown in Figure 3. A regular multidimensional association mining of data warehouse starts from the query formulation by a user. The mining engine running in a passive way based on the query the user has defined and the resulting rules are sent back to the users. The successful mining query and its user profile are stored in the mining log spontaneously. The system's query formulation interface provides browse and recommendation functions by the support of the user preference ontology, which also supports the active mining mechanism that we proposed here.

In [6], the approach of event-condition-action toward active database systems is very common. The similar way can be applied to active data mining. An active multidimensional association mining initiates a mining process automatically by the system while a

certain event take place. The events are classified into system detected and user defined. The system detected events are mostly motivated by the change of data warehouse while the user defined events are specific to individual user's inquiring.

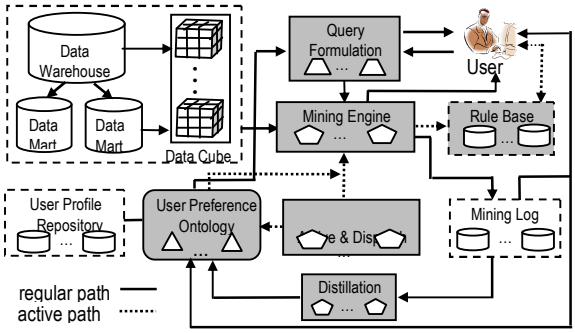

Figure 3. The active mining system framework

The active mining will trigger a macro or a micro action. The macro action launches a mining process that will run through all the mining queries maintained in the user preference ontology while the micro action will involve only certain part of the queries in the user preference ontology. Table 1 shows some examples.

Table 1. Examples of active events and actions

Event	Action	Type
new monthly data into DW	macro	system detected
new items > threshold	micro	system detected
user subscription	micro	user defined
renewal of user preference ontology	macro	system detected

The mining results will be stored in the rule base with temporal information which provides with resources for analyzing the long term rule trends. The user defined events can be specific time settings or period time settings for re-mining a specific query. The system dispatches specific rules that are newly generated to the users according to their preference, which is identified by the associations of the user profile index to the surrogate queries in the user preference ontology.

4. Related work

As surveyed in [6], considerable proposals and applications have been provided to the active database systems. Yet not much research on the active data mining was given. Data mining under changing environment is an important issue first introduced by Agrawal and Psaila [1]. They presented active data mining from accumulated association rules when certain trend of rules is found. They focus on defining shapes of trends and the shape query languages. Some other researches focused on dealing with the changes of classification mining [3][5][6][7].

5. Conclusions

We have proposed in this paper an active multidimensional association mining system framework with the help of the user preference ontology. We describe how the user preference ontology is constructed and how the active system framework works. The user preference ontology is the key to the active mechanism we propose. With the assistance of the user preference ontology, the mining queries issued frequently before are actively re-issued by the system when defined events turn true. The system is able to dispatch the new rules to the users according to their preference. The active mining scheme is useful that the possible new knowledge can be delivered to the user dynamically without any user's involvement.

Acknowledgements

This work is partially supported by National Science Council of Taiwan under grant No. NSC95-2221-E-390-024.

References

[1] R. Agrawal and G. Psaila, "Active data mining," in *Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining*, pp. 3-8, 1995.

[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, pp. 27–34, 1996.

[3] V. Ganti, J. Gehrke, and R. Ramakrishnan, "A framework for measuring changes in data characteristics," in *Proc. 8th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 126-137, 1999.

[4] W.H. Inmon, *Building the Data Warehouse*, John Wiley & Sons, Inc., New York, NY, 1995.

[5] B. Liu, W. Hsu, H.S. Han, and Y. Xia, "Mining changes for real-life applications," in *Proc. 2nd Int. Conf. on Data Warehouse and Knowledge Discovery*, pp. 337-346, 2000.

[6] N.W. Paton, O. Diaz, "Active database systems," *ACM Computing Surveys* (CSUR), Vol. 31, No. 1 , pp. 63-103, 1999.

[7] K. Wang, S. Zhou, C.A. Fu, and J.X. Yu, "Mining changes of classification by correspondence tracing," in *Proc. SIAM Int. Conf. on Data Mining*, pp. 97-106, 2003.