Mining Association Rules between Sets of Items in Large Databases

A. KrishnaKumar, D. Amrita, N. Swathi Priya

Abstract- In Data Mining, the usefulness of association rules is strongly limited by the huge amount of delivered rules. To overcome this drawback, several methods were proposed in the literature such as item set concise representations, redundancy reduction, and post processing. However, being generally based on statistical information, most of these methods do not guarantee that the extracted rules are interesting for the user. Thus, it is crucial to help the decision-maker with an efficient post processing step in order to reduce the number of rules. This paper proposes a new interactive approach to prune and filter discovered rules. First, we propose to use ontologies in order to improve the integration of user knowledge in the post processing task. Second, we propose the Rule Schema formalism extending the specification language proposed by Liu et al. for user expectations. Furthermore, an interactive framework is designed to assist the user throughout the analyzing task. Applying our new approach over voluminous sets of rules, we were able, by integrating domain expert knowledge in the post processing step, to reduce the number of rules to several dozens or less. Moreover, the quality of the filtered rules was validated by the domain expert at various points in the interactive process.

Keywords- Clustering, classification, and association rules, interactive data exploration and discovery, knowledge management applications.

I. INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.

Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Although data mining is a relatively new term, the technology is not.However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.

Association discovery in databases. Among sets of items in transaction databases, it aims at discovering implicative tendencies that can be valuable information for the decision-maker. An association rule is defined as the implication X → Y, described by two interestingness measures—support and confidence—where X and Y are the sets of items and $X \cap Y = Ø$.

Apriori is the first algorithm proposed in the association rule mining field and many other algorithms were derived from it. Starting from a database, it proposes to extract all association rules satisfying minimum thresholds of support and confidence. It is very well known that mining algorithms can discover a prohibitive amount of association rules; for instance, thousands of rules are extracted from a database of several dozens of attributes and several hundreds of transactions. Valuable information is often represented by those rare—low support—and unexpected association rules which are surprising to the user. So, the more we increase the support threshold, the more efficient the algorithms are and the more the discovered rules are obvious, and hence, the less they are interesting for the user. As a result, it is necessary to bring the support threshold low enough in order to extract valuable information. Rule mining, introduced in, is considered as one of the most important tasks in Knowledge.

Experiments show that rules become almost impossible to use when the number of rules overpasses 100. Thus, it is crucial to help the decision-maker with an efficient technique for reducing the number of rules. To overcome this drawback, several methods were proposed in the literature. On the one hand, different algorithms were introduced to reduce the number of item sets by generating closed , maximal or optimal item sets , and several algorithms to reduce the number of rules, using non redundant rules , or pruning techniques . On the other hand, post processing methods can improve the selection of discovered rules. Different complementary post processing methods may be used, like pruning, summarizing, grouping, or visualization. Pruning consists in removing uninteresting or redundant rules. In summarizing, concise sets of rules are generated. Groups of rules are produced in the grouping process; and the visualization improves the readability of a large number of rules by using adapted graphical representations. However, most of the existing post processing methods are generally based on statistical information in the database. Since rule interestingness strongly depends on user knowledge and goals, these methods do not guarantee that interesting rules will be extracted. For instance, if the user looks for unexpected rules, all the already known rules should be pruned. Or, if the user wants to focus on specific schemas of rules, only this subset of rules should be selected. Moreover, as suggested in, the rule post processing methods should be imperatively based on a strong interactivity with the user.

A. ONTOLOGIES IN DATA MINING

Data mining describes the association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases.

Manuscript received April, 2013.

N.SwathiPriya, Department Of Information Technology, SNS College Of Engineering, Coimbatore (TN), India.

A.KrishnaKumar, Department Of Information Technology, SNS College Of Engineering, Coimbatore (TN), India.

D.Amrita, Department Of Information Technology, SNS College Of Engineering, Coimbatore (TN), India.

It is intend to identify strong rules discovered in the databases using different measures of interestingness association rule learning is a popular and well researched method for discovering interesting relations between variables in the large databases. Ontologies have evolved over the years from controlled vocabularies to thesauri (glossaries), and later, to taxonomies.. By conceptualization, we understand here an abstract model of some phenomenon described by its important concepts. The formal notation denotes the idea that machines should be able to interpret ontology. Moreover, explicit refers to the transparent definition of the ontology elements. Finally, shared outlines that ontology brings together some knowledge common to a certain group, and not individual knowledge. Several other definitions are proposed in the literature. For instance, ontology is viewed as a logical theory. The ontologies are described as (Meta) data schemas, providing a controlled vocabulary of concepts, each with an explicitly defined and machine processable semantics Depending on the granularity, four types of ontologies are proposed in the literature: upper (or top level) ontologies, domain ontologies, task ontologies, and application ontologies.

Top-level ontologies deal with general concepts; while the other three types deal with domain specific concepts. Ontologies, introduced in data mining for the first time in early 2000, can be used in several ways. Domain and Background Knowledge Ontologies, Ontologies for Data Mining Process, or Metadata Ontologies. Background Knowledge Ontologies organize domain knowledge and play important roles at several levels of the knowledge discovery process. Ontologies for Data Mining Process codify mining process description and choose the most appropriate task according to the given problem; while Metadata Ontologies describe the construction process of items. In this paper, we focus on Domain and Background Knowledge Ontologies. The first idea of using Domain Ontologies was introduced by Srikant and Agrawal with the concept of Generalized Association Rules (GAR).

Ontology of background knowledge can benefit all the phases of a KDD cycle described in CRISP-DM. The role of ontologies is based on the given mining task and method, and on data characteristics. From business understanding to deployment, the authors delivered a complete example of using ontologies in a cardiovascular risk domain. Related to Generalized Association Rules, the notion of raising was presented. Raising is the operation of generalizing rules (making rules more abstract) in order to increase support in keeping confidence high enough. This allows for strong rules to be discovered and also to obtain sufficient support for rules that, before raising, would not have minimum support due to the particular items they referred to. The difference with Generalized Association Rules is that this solution proposes to use a specific level for raising and mining.

The knowledge discovery and data mining (KDD) field draws on findings from statistics, databases, and artificial intelligence to construct tools that let users gain insight from massive data sets. People in business, science, medicine, academia, and government collect such data sets, and several commercial packages now offer general-purpose KDD tools. An important KDD goal is to "turn data into knowledge." For example, knowledge acquired through such methods on a medical database could be published in a medical journal. Knowledge acquired from analyzing a financial or marketing database could revise business practice and influence a management school's curriculum. In addition, some US laws require reasons for rejecting a loan application, which knowledge from the KDD could provide. Occasionally, however, you must explain the learned decision criteria to a court, as in the recent lawsuit Blue Mountain filed against Microsoft for a mail filter that classified electronic greeting cards as spam mail. In one early KDD success story, Robert Evans and Doug Fisher analyzed data from a printing press, found conditions under which the press failed, and identified rules to avoid these failures. Unfortunately, for every insightful nugget we find, there are many more obvious or trivial rules Perhaps more troubling is that some rules are counterintuitive. For example, in screening for Alzheimer's disease, we found the following counterintuitive rule: "If the years of education of the patient are greater than the patient does not know the date and the patient does not know the name of a nearby street, then the patient is normal."

The search strategy of our algorithm integrates a depth-first traversal of the item set lattice with effective pruning mechanisms. Our implementation of the search strategy combines a vertical bitmap representation of the database with an efficient relative bitmap compression schema. In a thorough experimental analysis of our algorithm on real data, we isolate the effect of the individual components of the algorithm.

II. LITERATURE SURVEY

U.M. Fayyad et al. proposed a order to discover only those association rules that are interesting according to these measures. They have been divided into objective measures and subjective measures. Objective measures depend only on data structure. Many survey papers summarize and compare the objective measure definitions and properties. Unfortunately, being restricted to data evaluation, the objective measures are not sufficient to reduce the number of extracted rules and to capture the interesting ones. Several approaches integrating user knowledge have been proposed. In addition, subjective measures were proposed to integrate explicitly the decision-maker knowledge and to offer a better selection of interesting association rules.

A. Silbershatz et al. proposed a classification of subjective measures in unexpectedness a pattern is interesting if it is surprising to the user and action ability a pattern is interesting if it can help the user take some actions. As early as 1994, in the KEFIR system, the key finding and deviation notions were suggested. Grouped in findings, deviations represent the difference between the actual and the expected values. KEFIR defines interestingness of a key finding in terms of the estimated benefits, and potential savings of taking corrective actions that restore the deviation back to its expected value. These corrective actions are specified in advance by the domain expert for various classes of deviations.

M.J. Zaki et al. proposed templates to describe the form of interesting rules (inclusive templates) and not interesting rules (restrictive templates).

The idea of using templates for association rule extraction was reused in other approaches proposed to use a rule-like formalism to express user expectations and the discovered association rules are pruned/summarized by comparing them to user expectations.

R. Agrawal et al. proposed a query language for association rule pruning based on SQL, called M-SQL. It allows imposing constraints on the condition and/or the consequent of the association rules.

J. Li et al. proposed architecture for query-based association rule pruning, but more constraints had driven exploratory mining of rules. The lack of user exploration and control, the rigid notation of relationship, and the lack of focus.

E. Baralis et al. proposed a new query language called Constrained Association Query and they pointed out the importance of user feedback and user flexibility in choosing interestingness metrics. Another related approach was proposed by An et al. in where the authors introduced domain knowledge in order to prune and summarize discovered rules. The first algorithm uses data taxonomy, defined by user, in order to describe the semantic distance between rules, and in order to group the rules.

D. Burdick et al. proposed algorithm allows grouping the discovered rules that share at least one item in the antecedent and the consequent. In 2007, a new methodology was proposed in to prune and organize rules with the same consequent. The authors suggested transforming the database in an association rule base in order to extract second-level association rules called meta rules, the extracted rules r1! r2 express relations between the two association rules and help pruning/grouping discovered rules.

III.PROPOSED SYSTEM

Consider a supermarket with a large collection of items. Typical business decisions that the management of the supermarket has to make include what to put on sale, how to design coupons, how to place merchandise on shelves in order to maximize the prot, etc. Analysis of past transaction data is a commonly used approach in order to improve the quality of such decisions. Until recently, however, only global data about the cumulative sales during some time period (a day, a week, a month, etc.) was available on the computer. Progress in bar-code technology has made it possible to store the so called basket data that stores items purchased on a per-transaction basis. Basket data type transactions do not necessarily consist of items bought together at the same point of time. It may consist of items bought by a customer over a period of time. Examples include monthly purchases by members of a book club or a music club. Current address: Computer Science Department, Rutgers University, New Brunswick, NJ 08903 Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery.

To copy otherwise, or to republish, requires a fee and/or special permission. Proceedings of the 1993 ACM SIGMOD Conference Washington DC, USA, May 1993 several organizations have collected massive amounts of such data. These data sets are usually stored on tertiary storage and are very slowly migrating to database systems. One of the main reasons for the limited success of database systems in this area is that current database systems do not provide necessary functionality for a user interested in taking advantage of this information. This paper introduces the problem of \mining" a large collection of basket data type transactions for association rules between sets of items with some minimum species condense, and presents and client algorithm for this purpose. An example of such an association rule is the statement that 90% of transactions that purchase bread and butter also purchase milk.

Starting from a database, it proposes to extract all set of association rules satisfying minimum thresholds of support and confidence on ontology. It is very well known that mining algorithms can discover a prohibitive amount of association rules; for instance, thousands of rules are extracted from a database of several dozens of attributes and several hundreds of transactions, valuable information are often represented by those rare, low support, and unexpected association rules which are surprising to the user. So,we increase the support threshold, the more efficient the algorithms are and the more the discovered rules are obvious, and hence, the less they are interesting for the user.It is necessary to bring the support threshold low enough in order to extract valuable information. Unfortunately, the lower the support is, the larger the volume of rules becomes, making it intractable for a decision-maker to analyze the mining result. Experiments show that rules become almost impossible to use when the number of rules overpasses 100. Thus, it is crucial to help the decision-maker with an efficient technique for reducing the number of rules. The Demerits of existing system are :

- Usefulness of association rules is strongly limited by the huge amount of delivered rules.
- It is crucial to help the decision-maker with an efficient technique for reducing the number of rules.

Toward the use of ontology concepts. Furthermore, an interactive and iterative framework is designed to assist the user throughout the analyzing task. The interactivity of our approach relies on a set of rule mining operators defined over the Rule Schemas in order to describe the actions that the user can perform. This project is structured as follows: introduces notations and definitions used throughout the paper. Here we proposed our motivations for using ontologies. Describes the research domain and reviews related works and its Presents the proposed framework and its elements. It is devoted to the results obtained by applying our method over a questionnaire database and finally we present conclusions and shows directions for future research.

The Merits of proposed System are:

- Reduce the number of item sets by generating closed, maximal optimal item sets, and several algorithms to reduce the number of rules, using nonredundant rules, and pruning techniques.
- Domain ontologies improve the integration of user domain knowledge concerning the database field in the post processing step.

- The integration of domain expert knowledge in the post processing step in order to reduce the number of rules to several dozens or less.

V. CONCLUSION

On the one hand, domain ontologies improve the integration of user domain knowledge concerning the database field in the postprocessing step. On the other hand, we propose a new formalism, called Rule Schemas, extending the specification language. The latter is especially used to express the user expectations and goals concerning the discovered rules.

REFERENCES

1. R. Agrawal, T. Imielinski, and A. Swami, “Mining Association Rules between Sets of Items in Large Databases,” Proc. ACM SIGMOD, pp. 207-216, 1993.
2. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.
3. A. Silberschatz and A. Tuzhilin, “What Makes Patterns Interesting in Knowledge Discovery Systems,” IEEE Trans. Knowledge and Data Eng. vol. 8, no. 6, pp. 970-974, Dec. 1996.
4. M.J. Zaki and M. Ogihara, “Theoretical Foundations of Association Rules,” Proc. Workshop Research Issues in Data Mining and Knowledge Discovery (DMKD ’98), pp. 1-8, June 1998.
5. D. Burdick, M. Calimlim, J. Flannick, J. Gehrke, and T. Yiu, “Mafia: A Maximal Frequent Itemset Algorithm,” IEEE Trans.Knowledge and Data Eng., vol. 17, no. 11, pp. 1490-1504, Nov. 2005.
6. J. Li, “On Optimal Rule Discovery,” IEEE Trans. Knowledge and Data Eng., vol. 18, no. 4, pp. 460-471, Apr. 2006.
7. M.J. Zaki, “Generating Non-Redundant Association Rules,” Proc. Int’l Conf. Knowledge Discovery and Data Mining, pp. 34-43, 2000.
8. E. Baralis and G. Psaila, “Designing Templates for Mining Association Rules,” J. Intelligent Information Systems, vol. 9, pp. 7-32, 1997.

AUTHORS PROFILE

Mr. A.Krishnakumar, received his B.Tech degree in Information Technology from Jayamatha Engineering ollege, Kaniyakumari in 2011 and presently purusing M.Tech degree in Information Technology, Coimbatore. His research interest Data Mining,Wireless Networking, and Mobile Computing.

Ms. D.Amrita, received her B.Tech degree in Information Technology from KTVR knowledge park for engineering and technology, Coimbatore in 2012 and presently purusing M.Tech degree in Information Technology, Coimbatore. Her research interest Data Mining,Wireless Networking.

Asst Prof. N.Swathi Priya, presently working as a assistant professor in SNS College of Engineering-Coimbatore. Her research interest Data Mining,Wireless Networking.