ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

A Knowledge Based Post Mining Techniques in Large Databases through Interactive Post Processing of Association Rules using Ontologies

A. Vaishnavi[1], M. Hemalatha[2]

Assistant professor, Department of Computer Applications, Pioneer College of Arts and Science[1,2]

Abstract: In Data Mining, Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. the usefulness of association rules is vigorous limited by the huge amount of delivered rules. To overcome this drawback, several methods were proposed in the literature such as itemset concise representations, redundancy reduction, and postprocessing. Although, being generally based on statistical report, most of these methods do not guarantee that the extracted rules are interesting for the user. Thus, it is critical to help the decision-maker with an efficient postprocessing step in order to reduce the number of rules. This paper proposes a new interactive approach to prune and filter discovered rules. First, it propose to use ontologies in order to improve the integration of user knowledge in the postprocessing task. Second, it proposes the Rule internal representation of formalism extending the specification language proposed by Liu et al. for user expectations. Third it proposes to use the same in large databases for an effective and efficient result with out loss of an interesting item set. This paper system will reduce the number of rules with out loss an interesting item set while dealing with Large Databases.

Keywords: Association rules, classification, interactive data exploration and discovery , Post processing Clustering

I. INTRODUCTION

Association rule mining, introduced in , is considered as one of the most important tasks in Knowledge Discovery in Databases [2].more sets of items in transaction databases, it aims at discovering implication tendencies that can be valuable information for the decision-maker. An association rule is defined as the implication X ! Y , described by two exciting computes[12]—support and confidence—where X and Y are the sets of items and X \ Y ¼ ;. Apriori is the one of the algorithm proposed in the association rule mining field and many other algorithms were derived from it. Starting from a database, it suggest 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 occurance, thousands of rules are extracted from a database of several dozens of attributes and several hundreds of transactions. Furthermore, as suggested by Silbershatz and Tuzilin , valuable information is often represented by those rare—low support—and unpredeicted association rules which are surprising to the user. So, the more increase the support threshold, the more efficient the algorithms are and the more the discovered rules are perceived, and hence, the less they are fcinating for the user. As a result, it is more important to bring the support threshold low enough in order to extract valuable information. From the perspective of many types of practical decision aiding applications, however, both data mining and decision analysis techniques have some disadvantages.In decision support system development, there is little effort for generating synergies with enhancing each other's restrictions. More specifically, user bias, which play a key role in decision assists with decision analysis, are not definite considered in the contemporary generation of data mining systems. Even if they are (indirectly) aim at, they are constrained to the partiality of the data mining engineers by the use of threshold values rather than the decision makers' preferences that should be extend and adjusted to the current dynamic business environment. Decision analysis is not compatible with extracting knowledge from large corporate databases of nowadays, consideration of it does not focus on the automotive generation of meaningful knowledge from raw data post processing methods can improve the selection of discovered rules. Different interdependent post processing methods may be used, like pruning, summarizing, grouping, or visualization. Pruning consists in removing uninteresting or redundant rules. In summarizing, incisive sets of rules are generated. Categories of rules are produced in the grouping process; and the visualization exceeds the legibility of a large number of rules by using adapted graphical representations.

II .DEFINITIONS

2.1 Association Rule Mining

Association rule mining searches for interesting relationships more items in a given data set.

2.2 Associations and Item-sets

An association is a rule of the form: if X then Y, It is denoted as X → Y

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

Example: If India wins in cricket, sales of sweets go up.

2.3 Interesting item-set

For any rule if X → Y ➔ Y → X, then X and Y are called an "interesting item-set".

Example: People buying school uniforms in June also buy school bags

Association Rule Mining Factors

A rule X ⇛ Y is described using two important statistical factors: Support and Confidence**.**

Support (%)

Fraction transaction that both X and Y.

Confidence-(strength of the rule)

Measure how often items in Y appears in transactions that contain x

$$Support\ S: = \frac{\sigma(X \text{ and } Y)}{|T|}$$

$$Confidence\ C: = \frac{\sigma(X \text{ and } Y)}{\sigma(X)}$$

Some important definitions as follows were used with the references that are Transactions that contain the itemset, An association rule[7] is an implication, Maximal itemset, Galois closure operators, A closed itemset , rules having minimal antecedents and ensuing , in terms of subset relation. A rule set is optimal , an ontology[16] is a quintuple.

III. EXISTING SYSTEM

The existing system is composed of two main parts First, the knowledge base allows formalizing user understanding and objectives. Domain knowledge offers a general view over user knowledge in database domain, and user assumptions express the prior user knowledge over the recognized rules. Second, the post processing task consists in applying iteratively a set of filters over extraction rules in order to extract rules: minimum improvement constraint filter, item-relatedness filter, rule schema filters/pruning.

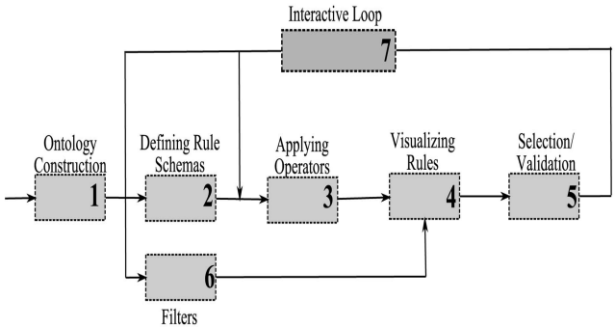


Fig. 1. Interactive process description

The novelty of this approach resides in supervising the knowledge discovery process using two different ideal structures for user knowledge representation: one or several ontologies[16] and several rule schemas generalizing general impressions, and proposing an iterative process.

The ARIPSO framework proposes to the user an synergistic.process of rule discovery, presented in "Fig. 1." Taking into account his/her feedbacks, the user is skillful to revise his/her expectations in function of intermediate results. Several steps are suggested to the user in the framework as follows:

1. ontology construction—starting from the database, and finally, from existing ontologies, the user develops an ontology on database items;

2. defining Rule Schemas (as GIs and RPCs)—the user expresses his/her local goals and expectations concerning the association rules that he/she wants to find;

3.To pick the right operators to be applied over the rule schemas created, and then, applying the operators;

4. visualizing the results—the filtered rules forward to the user;

5. selection/validation—starting from these preliminary results, the user can validate the results or he/she can revise his/her information;

6. This system has user two filters already existing in the literature. These two filters can be appeal over rules whenever the user needs them with the main goal of reducing the number of rules; and

7. the interactive loop allows to the user to revise the information that he/she proposed. Thus, he/she can return to step 2 in order to make the modification of the rule schemas, or he/she can return to step 3 in order to change the operators. Besides, in the interactive loop, the user could decide to apply one of the two predefined filters discussed in step 6

This system author used, to filter four types of rules using: keep rules and unexpected rules concerning the antecedent and/or the consequent:

. Conforming rules—association rules that are conforming to the define beliefs;

. Unexpected antecedent rules—association rules that are unexpected regarding the antecedent of the specified beliefs;

. Unexpected consequent rules—association rules that are unexpected regarding the consequent of the specified beliefs; and

. Both side unexpected rules—association rules that are unexpected regarding both the antecedent and the consequent of the specified beliefs.

The following "Fig.'s" will give the better understanding of ontology for a Supermarket item taxonomy example

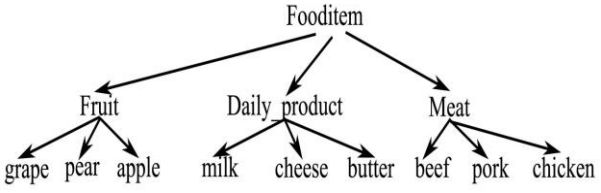


Fig.2. Supermarket item taxonomy

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

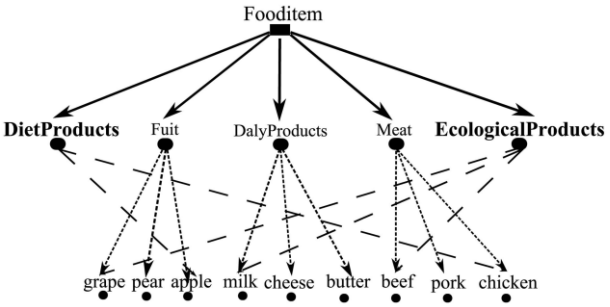


Fig.3. Visualization of the ontology created based on the supermarket item taxonomy

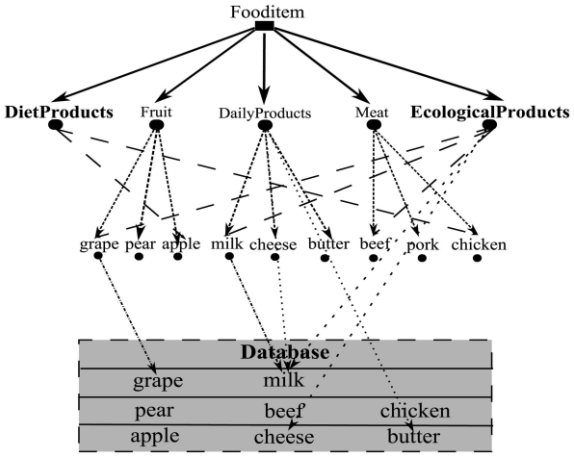


Fig. 4. Ontology description

IV. PROPOSED SYSTEM

It propose two important operators: pruning and filtering operators. It has three operations conforming, unexpectedness, and exception. Operators in the postprocessing task: are pruning and exceptions. To reduce the number of rules the filter were used The item-relatedness filter (IRF) was proposed by Shekar and Natarajan[26] . users are interested to find association between itemsets[5] with different functionalities, coming from different domains. So it use integrated filters to get an effective result[21]. In large data bases the number of transaction and conditions were more here. Improved Apriori algorithm is used to mine association rules support was identified by number of transaction and type of compliant registered and confidence will be generated by Ontologies description. Conceptual Structure of the Ontology and Ontology-Database Mapping were used to finalize the description.

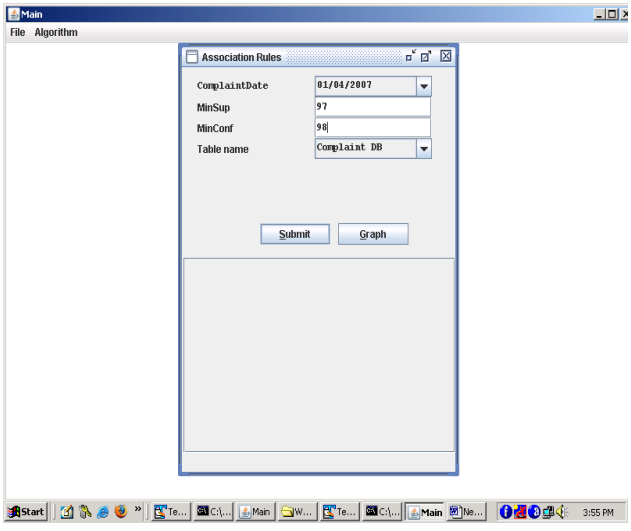
This system also has an apriori algorithm but the efficiency of an algorithm is improved by

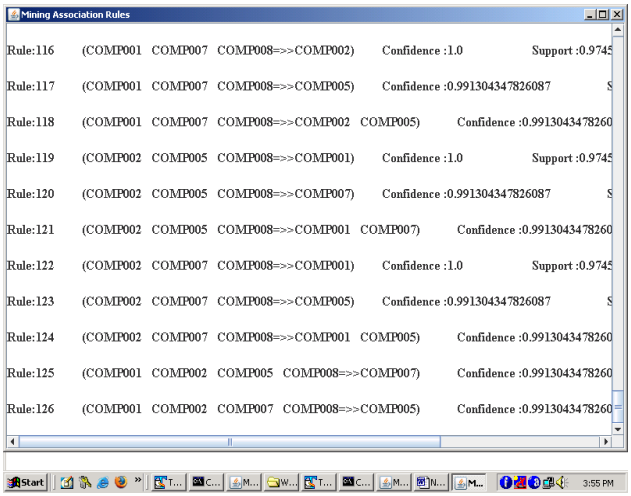
- Dynamic item set counting: add new candidate item sets only when all of their subsets are estimated to be frequent.
- Sampling: mining on a subset of given data, lower support threshold + a method to determine the completeness.
- Partitioning: Any itemset that is potentially frequent in DB must be frequent in at least one of the partitions of DB.[5].
- Transaction reduction: A transaction that does not contain any frequent k-itemset is useless in subsequent scans.
- Hash-based itemset counting: A k-itemset whose corresponding hashing bucket count is below the threshold cannot be frequent.

Dynamic item set counting is mainly concentrated with minimum threshold with an minimum support and confidence. Above consideration for a large base.

V. CONCLUSION

This paper discusses the problem of selecting interesting association rules all around huge volumes of discovered rules. The improved apriori algorithm is used to find the recurrent item set, it Reduce the passes of transaction database scans. It allows integration of domain expert knowledge in the postprocessing. It will reduce the number of rules without loss of an interesting item set while having large databases.





ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases,"
[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.
[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.,
[6] J. Li, "On Optimal Rule Discovery," IEEE Trans. Knowledge and Data Eng.,
[7] M.J. Zaki, "Generating Non-Redundant Association Rules," Proc. Int'l Conf. Knowledge Discovery and Data Mining, pp. 34-43, 2000.
[8] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Efficient Mining of Association Rules Using Closed Itemset Lattices,"
[9] H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hatonen, and H. Mannila, "Pruning and Grouping of Discovered Association Rules," Proc. ECML-95 Workshop Statistics, Machine Learning, and Knowledge Discovery in Databases, pp. 47-52, 1995.
[10] B. Baesens, S. Viaene, and J. Vanthienen, "Post-Processing of Association Rules," Proc. Workshop Post-Processing in Machine Learning and Data Mining: Interpretation, Visualization, Integration, and Related Topics with Sixth ACM SIGKDD, pp. 20-23, 2000.
[11] J. Blanchard, F. Guillet, and H. Briand, "A User-Driven and Quality-Oriented Visualization for Mining Association Rules," Proc. Third IEEE Int'l Conf. Data Mining, pp. 493-496, 2003.
[12] B. Liu, W. Hsu, K. Wang, and S. Chen, "Visually Aided Exploration of Interesting Association Rules," Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD), pp. 380-389, 1999.
[13] G. Birkhoff, Lattice Theory, vol. 25. Am.
[14] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering Frequent Closed Itemsets for Association Rules," Proc. Seventh Int'l Conf. Database Theory (ICDT '99), pp. 398-416, 1999.
[15] M. Zaki, "Mining Non-Redundant Association Rules," Data Mining and Knowledge Discovery,
[16] A. Maedche and S. Staab, "Ontology Learning for the Semantic Web," IEEE Intelligent Systems,
[17] B. Liu, W. Hsu, L.-F. Mun, and H.-Y. Lee, "Finding Interesting Patterns Using User Expectations," .
[18] I. Horrocks and P.F. Patel-Schneider, "Reducing owl Entailment to Description Logic Satisfiability," .
[19] J. Pei, J. Han, and R. Mao, "Closet: An Efficient Algorithm for Mining Frequent Closed Itemsets," Proc. ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery, pp. 21-30, 2000.
[20] M.J. Zaki and C.J. Hsiao, "Charm: An Efficient Algorithm for Closed Itemset Mining," Proc. Second SIAM Int'l Conf. Data Mining, pp. 34-43, 2002.
[21] M.Z. Ashrafi, D. Taniar, and K. Smith, "Redundant Association Rules Reduction Techniques," AI 2005: Advances in Artificial Intelligence – Proc 18th Australian Joint Conf. Artificial Intelligence pp. 254-263, 2005.
[22] M. Hahsler, C. Buchta, and K. Hornik, "Selective Association Rule Generation," Computational Statistic,
[23] J. Bayardo, J. Roberto, and R. Agrawal, "Mining the Most Interesting Rules," Proc. ACM SIGKDD, pp. 145-154, 1999.
[24] R.J. Bayardo, Jr., R. Agrawal, and D. Gunopulos, "Constraint-Based Rule Mining in Large, Dense Databases,"
[25] E.R. Omiecinski, "Alternative Interest Measures for Mining Associations in Databases".
[26] R. Natarajan and B. Shekar, "A Relatedness-Based Data-Driven Approach to Determination of Interestingness of Association Rules,"