Interactive Approach for Generation of Association Rules by using Ontology

Mr.Rahul Jadhav
(Assistant Professor)
Fr. C. Rodrigues Institute of Technology,
Agnel Technical Education Complex,
Sector-9A, Vashi,
Navi Mumbai- 400703.

***Abstract*—Data mining field consist of many operations on data.Association rule generation is one of the very important operations. Association rule mining finds interesting association or correlation relationships among a large set of data items. Methods proposed to generate association rules are mainly based on statistics .These methods generatea huge voluminous number of association rules. many times, user gets rules in which he is not at all interested, so dependency on statistical methods are not sufficient .Paper proposes new approach which integrate user knowledge in interactive framework through ontology. The rules generated through statistical methods such as Apriory or frequent pattern mining algorithm undergoes analyzing task and permit user to review provided rules. Ontology process guarantees that the generated rules are as per user's interest. Statistical filtering methods are used to prune and filter the association rules .Output of filtering process provide optimal rules as per users expectations and need.**

***IndexTerms*—Data Mining,Association Rule,Ontology**

INTRODUCTION

Association rule mining finds interesting association or correlation relationships among a large set of data items. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining association rules from the databases. The discovery of interesting association relationships among huge amounts of business transaction records can help in business decision making processes, such as catalog design, cross-marketing, and loss-leader analysis. The market basket analysis process analyzes customer buying habits by finding associations between different items that customers place in their shopping baskets.

This paper proposes a new interactive approach to prune and filter discovered rules. Ontologies will be used to improve the integration of user knowledge in the post processing task. Secondly, the Rules Schema formalism is proposed for user expectations. Hence when new approach applied to voluminous sets of rules, integrating domain expert knowledge in the post processing step, will reduce the number of rules to several dozens or less, these rules are strong rules in which user is interested in.

A number of rules are delivered in association rule mining; hence the strong rules for decision making are lost. The main purpose is to reduce the rules using post processing techniques as to get the strong rules for a good decision making.

Some association rule mining Definitions are:

An Association rule is an implication X→ Y, where X and Y are two itemsets and X intersection Y =0. The X, is called the antecedent of the rule, and the later, Y, is called the consequent.

Support: The support of the rule, that is, the relative frequency of transactions …The Support of the rule, defined as,

supp(X → Y) = supp (X U Y) = | t(X U Y)|, is the ratio of the number of transaction containing X U Y. If supp (X → Y) =s, s % of transactions contains the itemset X UY.

Confidence: The confidence of the rule, defined as, conf(X→Y)=supp(X→Y)/supp(X)=supp(XUY)supp(X)=c,is the ratio (c%) of the number of transactions that containing X, contain also Y.

Lift: The lift value of the rule is the additional interestingness measures on the rules. These measures can then be used to rank the rules by importance (and present a sorted list to the user) or as an additional pruning criterion.

Ontology: Ontology is a quintuple.

O ={C,R,I,H,A} ; C ={C1; C2; . . . ; Cn} is a set of concepts and R ={R1;R2; . . .Rm} is a set of relations defined overconcepts. I is a set of instances of concepts and H is a DirectedAcyclic Graph (DAG) defined by the sub assumption relation between concepts. We say that C2 is-a C1,C1 ≤ C2, if the concept C1 subsumes the concept C2.Ais a set ofaxioms bringing additional constraints on the ontology.

In association rule mining, frequent itemsets have to be deduced first and then the strong rules are obtained. Unfortunately, the lower the support count is, larger the volume of rules becomes, making it impractical for a decision-maker to analyze the mining result. Experiments show that rules become almost impossible to use when the number of rules overpass 100. Thus, it is crucial to help the decision-maker with an efficient technique for reducing the number of rules. There are number of existing approaches to reduce the frequent itemsets.

1) Apriori algorithm (Forward Scanning)

Apriori algorithm for finding Frequent (Large) Item sets:

The first algorithm to use the downward closure and antimonotonicity properties is the Apriori Algorithm. Candidate itemsets with one item are scanned throughout database to calculate their support. The candidate itemsets having support greater than min_sup forms frequent itemset with one item. Union of two frequent itemsets having n items produces candidate itemset with n+ 1 item then form frequent itemsets from these candidate itemsets till we get frequent itemset having maximum item.

Steps are involved in this:

A. Mining frequent itemset:

This will be done using a database and applying an FP tree growth algorithm to the database to generate frequent item-sets.Finding the set of items that have minimum support .A subset of a frequent itemset should also be frequent itemset, i.e if {AB} is a frequent itemset, both {A} and {B} should be frequent itemsets. Iteratively find frequent itemsets with cardinality 1 to k (k-itemsets).

B. Generating association rules from frequent itemsets : Procedure:

a) For each frequent itemset "I", generate all nonempty subsets of I.

b) For every Non empty subset of I, output the rule " s→(I-s)"

If support_count(I)/support _count(s)>=min_conf where min_conf is minimum confidence threshold.

2) Frequent Pattern (FP) tree Algorithm

This algorithm does not use candidate item sets and works as follows:

Start from each frequent length-1 pattern (as an initial suffix pattern). Construct its conditional pattern base. Construct its conditional FP-Tree& perform mining on such a tree. The pattern growth is achieved by concatenation of the suffix pattern with the frequent patterns generated from a conditional FP-Tree. The union of all frequent patterns gives the required frequent item set. FP tree is constructed using two passes

Pass 1:

Scan data and find support for each item.

That is Discard infrequent items after that Sort frequent items in decreasing order based on their support later Use this order when building the FP-Tree, so common prefixes can be shared

Pass 2:

Find Nodes correspond to items and have a counter.FP-Growth reads 1 transaction at a time and maps it to a pathFixed order is used, so paths can overlap when transactions share items (when they have the same prefix). In this case, counters are incremented Pointers are maintained between nodes containing the same item, creating singly linked lists (dotted lines)The more paths that overlap, the higher the compression. FP-tree may fit in memory.at this stage frequent itemsets extracted from the FP-Tree

EXISTING APPROACHES

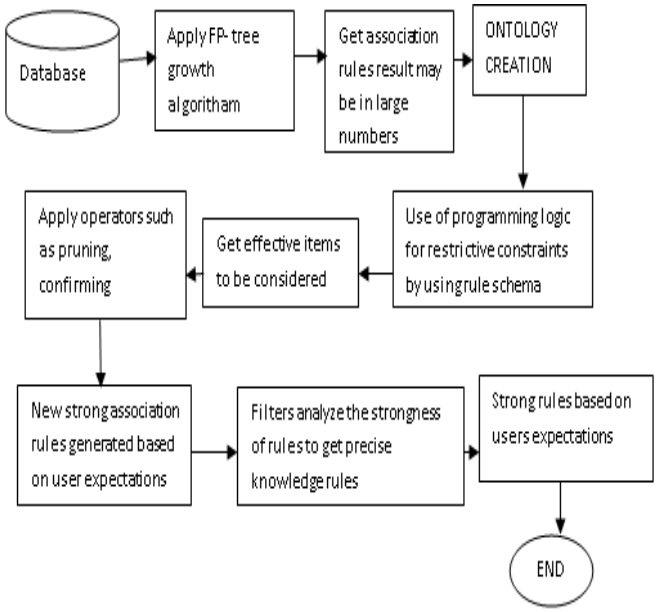
ARIPSO(Association Rule Interactive Post-Processing using Schemas and ontologies) Framework [1].Post processing task can be improved by using domain ontology. User's expectations are accepted through Rules schema. In this approach author has suggested to use ontology and various pruning operations basically approach goes in generation of association rules, gathering user's expectations and post processing [1].

A paper [3] has an approach for automatic generation of Personal Web Usage Ontology of periodic access patternsfrom web usage logs that have been semantically enrichedwith information on emotional influence and resourcetopics]. Propose a semantic webusage mining approach for discovering periodic web access patterns from annotated web usage logs which incorporates informationon consumer emotions and behaviors through self-reporting and behavioral tracking. We use fuzzy logic to represent real-life temporalconcepts (e.g., morning) and requested resource attributes (ontological domain concepts for the requested URLs) of periodic patternbasedweb access activities. [3]

In the paper [5], the authors haveproposed framework for generating the combined rules which gives informative knowledge for business by combining static and transactional data.

In the paper [4], the authorsprovidea new approach which gives pruning method to remove the redundant rules before generating the combined rules. These Rule Clusters are generated for similar group customer or similar transaction characteristics which provide more interesting knowledge and actionable result than traditional association rule. Experimental results demonstrate the proposed techniques.Rule pairs or rule clusters are done only after the pruning. Domain driven approach provides freedom to select the product

PROPOSED SYSTEM



Proposed system.

The entire system is divided into 3 main phases

1) Association Rule Mining
2) Ontologies
3) Post-processing Step

1) Association Rule Mining

Association rule mining searches for relationships between items in a dataset, i.e. finding association, correlation among sets of items or objects in transaction databases, relational databases and other information repositories.

Mining frequent itemset:

This will be done using a database and applying an FP tree growth algorithm to the database to generate frequent item-sets. Finding the set of items that have minimum support .A subset of a frequent itemset should also be frequent itemset, i.e if {AB} is a frequent itemset, both {A} and {B} should be frequent itemsets.Iteratively find frequent itemsets with cardinality 1 to k (k-itemsets).

2) Creating Ontology

Ontology is a term borrowed from philosophy that refers to the science of describing the kind of entities in the world and how they are related. The OWL Web Ontology Language is a language for defining and instantiating Web ontologies. The OWL Web Ontology Language is intended to provide a language that can be used to describe the classes and relation between them that are inherent in Web documents and applications. OWL language is used to formalize a domain by defining classes and properties of those classes, Define individuals and assert properties about them.Expressiveness of ontology is done by rule scemas.it combines abstaction and pruning constraints [2].

RS (<p1, p2… (→) q1, q2...>)

A Rule Schema expresses the fact that the userexpects certain elements to be associated in the extractedassociation rules.

WherePi; Qiare subset of Cwhere O is ontology, O={C, R, I, H, A} and the proposed formalism combines General Impressions andReasonably Precise Concepts.

3) Post-Processing Step

This is to reduce the number of rules, by applying certain constraints or filters. Comparing the association rules derived and the rules derived from user knowledge ontologies extract the final set of filtered rules.

Proposed system uses some operators for filtering. Pruning operator eliminates the association rules which are not matching with rule schemas for each generated rule, interesting measures like Confidence and Lift is calculated. Pruning selects only those rules which are not having both the properties dataset A as well as dataset B. In this step many of the rules get omitted as they are not showing properties of both dataset A and B. Rules matching with non-implicative rule schema are filtered using confirming operator. Statistical method such asMinimum improvement constraint filter(MICF)selects only those rules whose confidence is greater than the confidence of any of simplifications. [3]

E.g. Grape, pear→ milk (confidence=85%)

Grape → milk (confidence=90%)

Pear → milk (confidence=83%)

The rule Pear → milk is removed

Large database provides number of transactions. On such a database Apriory or Frequent Pattern growth algorithm can be applied to generate Association rules. Ontology is prepared using itemsets available in database. The user selects the items in which he is interested, the rules generated are based on those itemsets and rules other than user's interest are exempted. This can be done by using pruning operator. The generated rules after ontology are concise rules. On these rules filer is applied this again is statistical approach but ontology handles users interest so applying filters after ontology, filters out rules which doesn't get pass through filter. This process reduces rules by dozens or so and these will be strong association rules

CONCLUSION

The importance of data mining as a source of information has never been more than what it is today. One can only imagine it will grow in the future. Market Basket analysis is both the means and the facilitators to access this vast information in an efficient and user-friendly manner.

In this praposed system, association rule generation algorithm such as apriory or FP tree growth algorithm can be used for association rule mining to extract the strong rules from the datasets. However there are numerous rules which are deduced from the dataset and hence it is difficult to differentiate between the strong rules. One of the features in this approach is ontology, which enables optimization ofresult by specifying the rule schema as per the user requirement, thus allowing a controlled interaction between the user and the software, and deducing the strong rules for market basket analysis as per the requirement.the process can be iterative if user is not satisfied with generated rules after ontology .User can go for new item selection through ontology step.In filtering process unwanted rules are exempted through post processing step operaters are used to guide users

REFERENCES

1. Marinica, C., & Guillet, F. (2010). Knowledge-based interactive postmining of association rules using ontologies. Knowledge and Data Engineering, IEEE Transactions on, 22(6), 784-797
2. Sulthana, A. R., & Murugeswari, B. (2011, March). ARIPSO: Association rule interactive postmining Using Schemas And Ontologies. In Emerging Trends in Electrical and Computer Technology (ICETECT), 2011 International Conference on (pp. 941-946). IEEE.
3. Fong, A. C. M., Zhou, B., Hui, S., Tang, J., & Hong, G. (2012). Generation of personalized ontology based on consumer emotion and behavior analysis. Affective Computing, IEEE Transactions on, 3(2), 152-164.
4. Deshpande, A., Mahajan, A., Kulkarni, A., & Sakhalkar, S. (2013, August). Domain driven approach for coherent rule mining. In Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on (pp. 109-114). IEEE.
5. Jacinto, C., & Antunes, C. (2012, May). User-driven Ontology L0.earning from Structured Data. In Computer and InformationScience (ICIS), 2012 IEEE/ACIS 11th International Conference on (pp. 184-189). IEEE
6. Xiaoxin Yin, Jiawei Han, Jiong Yang, and Philip S. Yu, —Efficient Classification across Multiple Database Relations: A CrossMine Approach‖ , IEEE Trans On Knowledge And Data Engineering, Vol. 18, No. 6, June 2006. pp 770 -783