A Study on Post mining of Association Rules Targeting User Interest

P. Sarala[#1], S. Jayaprada[*2]

[#]*Department Of Computer Science and Engineering*
V.R.Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India
[*]*Department Of Computer Science and Engineering*
V.R.Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India

Abstract- **Association Rule Mining means discovering interesting patterns with in large databases. Association rules are used in many application areas such as market base analysis, web log analysis, protein substructures. Several post processing methods were developed to reduce the number of rules using nonredundant rules or pruning techniques such as pruning, summarizing, grouping or visualization based on statistical information in the database. As such, problem of identifying interest rules remind the same. Methods such as Rule deductive method, Stream Mill Miner (SMM), a DSMS (Data Stream Management Systems), Medoid clustering technique (PAM: Partitioning around medoids), Constraint-based Multi-level Association Rules with an ontology support were developed but are not effective. The number of rules generated by Apriori, FPgrowth depends on statistical measures such as support, confidence and may not suit the requirements of user. Methods that use ranking algorithm and IRF (Item Relatedness Filter) have the drawbacks of using filters during pruning stage. The paper studies methods that were proposed for post processing of association rules and proposes a new method for extracting association rules based on user interest using MIRO (Mining Interest Rules Using Ontologies) framework that uses correlation measures combined with domain ontology, succint constraints.**

Keywords- **Association Rules, Association Rule Mining, Ontology, correlation measures, user constraints.**

I. INTRODUCTION

One important area in data mining is concerned with the discovery of interesting association rules. Association Rules describes relationship within set of coexisting data. An association rule a → b implies the presence of the item set b when an item set a occurs in a database transaction. Apriori algorithm [14] extracts all association rules satisfying minimum thresholds of support and confidence. If the support threshold is low then we can extract more valuable information. But usually rules are high.

The number of patterns extracted by current data mining algorithm is huge and cannot be easily handled by end user. As a result more memory is required to store this huge set. Solutions such as frequent closed patterns, using filters, using redundancy rules, ontologies with semantics were proposed also compression on appropriate set of frequent patterns, frequent patterns asserting the interestingness of association rules which is evaluated by using relatedness based on relationship between itempair[13].

But these solutions doesnot aim at user interestingness, reason being rules extracted are due to statistical measures such as support and confidence. As such rules that are of user interest can be filter during iterations. User constraints can be categorized as succint constraints that can be pushed into the intial data selection process at the start of mining. Monotonic constraints can be checked and once satisfied not to do more constraint checking at their further pattern growth. Anti-monotonic constraints can pushed deep into the mining process to restrain pattern growth. Also instead of statistical measures we can use correlation measures such as lift, cosin and overall confidence.

Section 2 summarizes the existing methods that reduce the number of rules. Section 3 describes our proposed approach MIRO. Conclusions and future work are given in sections 4 and 5.

II. RELATED WORK

The post-processing of association rules were improved by using an approach to finding a minimum set of suitable objective measures. Objective measures are also called as representative measures. This approach finds the representative measures by using medoid clustering technique (PAM: Partitioning around medoids) .The technique [17] uses linear correlation index, in order to partition thirty-six interestingness measures into *k* clusters. The subset of representative measures on the studied data is then constituted by the *k* medoids obtained.

Advantages:
1. By applying this approach on a rule-based dataset consists of 123228 association rules, a set of thirty-six measures is classified into sixteen clusters.
2. This technique allows the user to obtain not only the suitable measures representing the hidden aspects in the dataset, but also a graphical representation to evaluate the clustering results.
Disadvantage:
Clustering by using linear correlation coefficient was not an efficient method than the clustering by using a distance calculation more adapted to the data.

Stream Mill Miner (SMM), a DSMS (Data Stream Management Systems) designed for mining applications [10]. This system was mainly focused on the problem of post-mining association rules generated from the frequent patterns detected in the data streams. This has been implemented by (i) the SWIM algorithm, which incorporates preferences and constraints to improve the search, (ii) a historical database of archived rules that supports summarization, clustering, and trend analysis, over these rules, (iii) a high level mining language to specify the mining process, and (iv) a powerful DSMS that can support the mining models and process specified in this language as continuous queries expressed in an extended SQL language with quality of service guarantees.

Advantages:
1. It provides a number of mining methods and operators that were fast and light enough to be used for online mining of massive and often bursty data streams.
2. The bulk of the candidate rules were filtered out by this system, so that only a few highly prioritized rules were sent to analyst for validation.
3. By using the *Rule Post-Miner (RPM)* module the analyst continuously monitors the system. By querying and revising the historical rule repository, he/she provides critical feedback to control and optimize the post-mining process.
Disadvantages:
1. SMM does not provide the visualization of the results of association rules mining.
2. It does not consist of efficient mining algorithms in the system, such as the Moment algorithm or the
CFI-stream algorithm than SWIM algorithm for finding frequent patterns from a stream.
3. It does not have the capability of running SMM in a distributed environment where multiple server parallely process data streams.

An integrated framework was developed for extracting constraint-based Multi-level Association Rules with an ontology support [1]. This method can improve the quality of filtered rules. There are number of ways to reduce the computational complexity of Association Rule Mining and number of ways to increase the quality of the extracted rules: (i) reducing the search space; (ii) exploiting efficient data structures; (iii) adopting domain-specific constraints. The first two classes of optimizations are used for reducing the number of steps of the algorithm, for re-organizing the itemsets, for encoding the items, and for organizing the transactions in order to minimize the algorithm time complexity. The third class tries to overcome the lack of user data-exploration by handling domain-specific constraints. This method mainly focuses on these optimizations by using ontology and expressing constraints in order to get the extracted association rules. The main purpose of this framework is to reduce the "search space" of the algorithm and to improve the significance of the association rules.

Advantages:
The main advantages of this framework were summarized in terms of extensibility and flexibility.
1) The framework was extensible because data properties and concepts can be introduced in the ontology without either changing the relational database containing the transaction, or the implementation of our framework.
2) The flexibility was guaranteed from the separation of the data to analyze (the transactions) from the meta-data (description of the data).
Disadvantages:
1) The overhead in conducting pruning tests it takes more time for execution.
2) This uses seRQL to express user knowledge which is not flexible than rule schema.

Closely related to our approach is the research on data mining from ontological data [6], in which they investigate on how to make current machine learning algorithms willing to work on instances that are described by means of an ontological vocabulary. They present a framework for designing kernels that exploit the knowledge of underlying ontologies. For their implementation, they extend the Support Vector Machine (SVM) algorithm with adequate kernels on Semantic Web data. The framework is based on common notions of similarity. They introduce four different kernels, residing on different layers. The identity layer kernel solely considers the identity of two instances, while the class layer considers similarities of instances based on the classes they instantiate. The property layers on the other hand, compare the similarities of instances based on the data properties and/or object properties. To evaluate their framework, the authors present two experiments on different datasets. Experiment one tries to imitate the classification behavior of an ontology given a semantically weakened ontology. In a second experiment they predict the research group affiliation of persons and publications based on the SWRC ontology.

A rule deductive method [15] was developed to mine the real demanded association rules for any given user. It provides friendly interface to the users and their needed association rules. It also provides dynamic response to users and let uses find their interested rules as soon as quick. This method does not need to scan the database for obtaining all the maximal frequent itemsets and avoids producing large number of frequent itemsets which consists of the upper long frequent itemsets. The dynamic response deductive strategies (deep search and wide search) can find most long frequent itemsets in short time. The deductive interesting rules mining method in the static situation effectively avoids producing huge amounts of frequent itemsets contained by the upper long frequent itemsets than the tranditional apriori algorithm. The dynamic response strategies of the deep search and wide search also efficiently find most long frequent itemsets in intial time and they could be applied in the online data mining.

Advantages:
1. This method avoids the mining of frequent itemsets starting from candidate two-itemsets to candidate (n-1)-itemsets.
2. It avoids the scanning of database for mining frequent pattern.
3. It avoids producing huge amounts of frequent itemsets.
4. It provides dynamic response to users in any time when users want to check whether their interested frequent itemsets have been founded or not.

Disadvantage:
It treats that providing dynamic response to users in any time was difficult.

A new approach was [7] developed to prune mined association rules in large databases and also provided different association rule mining techniques. This approach was consists of two main phases. The first phase includes the generation of support counts of item sets at each timeslot and candidate item sets. The second phase involves mining of association rules from candidate items and post mining of association rules using ontology and user constraint template to guarantee user interesting rules.

Advantages:
1) This approach was prune and filter the discovered rules.
2) It Guarantees the rules are interesting for the user.
3) The use of ontology provides specification of several characteristics of a domain.

Disadvantage:

This method takes more time for mapping ontology concepts with the DB items.

Link mining for the Semantic Web [8] introduces the importance of link mining for the Semantic Web and investigate the appropriate handling of correlations between entities. They state that the links among objects demonstrate certain patterns, which can be helpful for many data mining tasks and are usually hard to capture with traditional statistical models. Other studies [3,12] discuss the discovery of association rules in RDF data by introducing algorithms or operators.

The ARIPSO (Association Rule Interactive Postmining Using Schemas And Ontologies) [2] was developed to mine the interesting association rules from huge amount of rules. This approach used techniques such as Ontologies, Rule Schemas, interactive frame work, ranking approach and privacy in mining the rules. And it also used IRF filters inorder to avoid unwanted rules using different operators. These operators are used in the postprocessing task in mining and guide the user in the over all process.

Advantage: This ARIPSO framework was used for pruning and filtering the needed rules with respect to the expectation of the user.
Disadvantage: It requires some other filters to prune the rules.

Interesting measures such as directed information ratio based on information theory is designed by [5]. This rule will filter out the rules whose antecedent and consequent are negatively correlated [9] proposed a new significance assessment that not only depends on monotonic attributes but also on Chi –Square test [16] proposes a user feedback system that guides discovering of interesting patterns.

Methods such as [4] mines a compressed set of frequent patterns where as [14] summarizes a large set of patterns with the most representative once both works use extra information of frequent patterns beyond support but their work is restricted to morphological information or simple statistics and could not infer semantics within patterns.

III. PROPOSED APPROACH

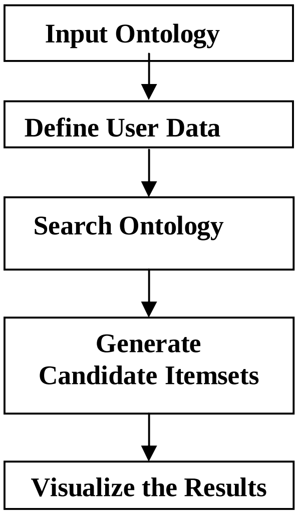


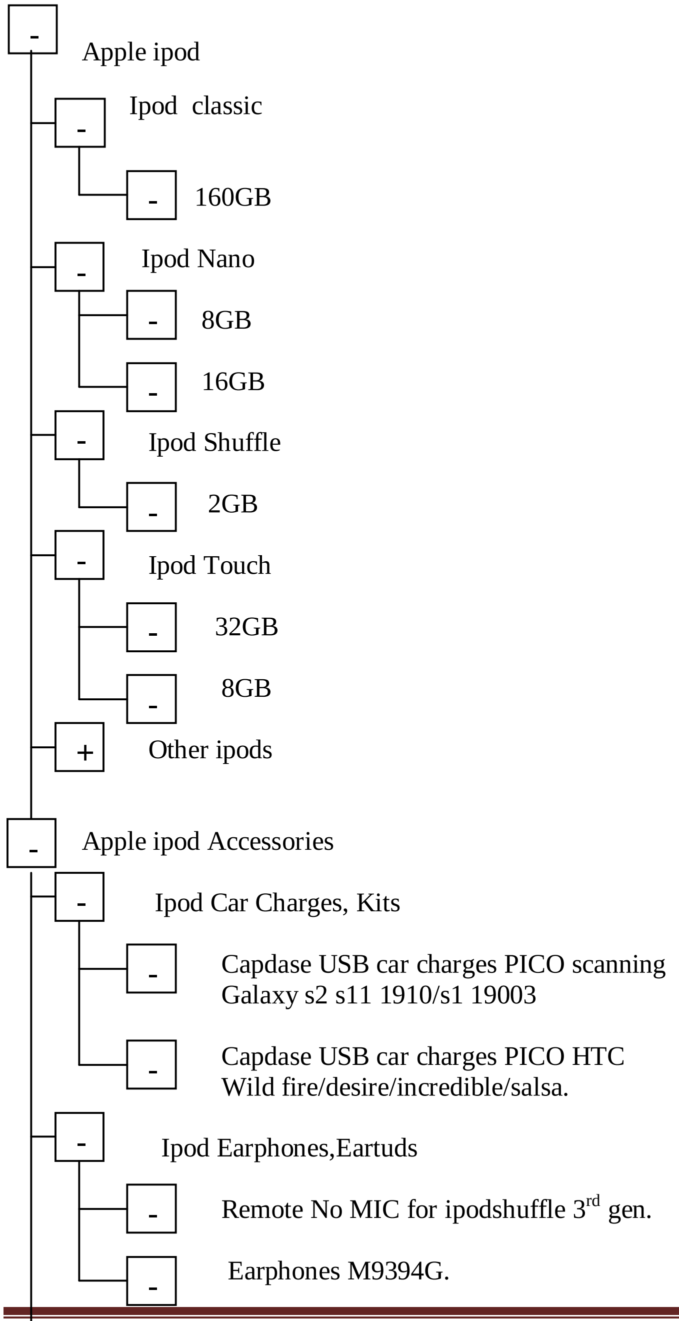
Figure 1: MIRO Process Description

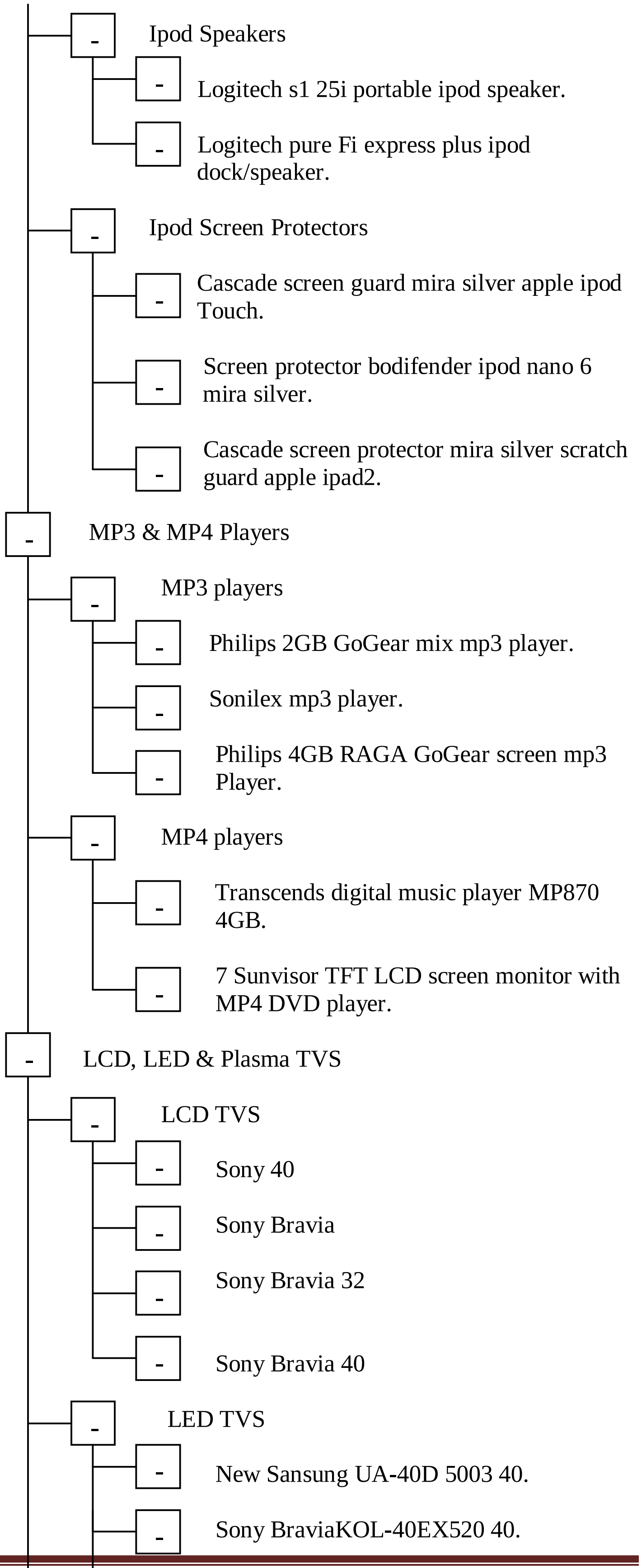
Figure 1 shows MIRO (Mining Interest Rules Using Ontologies) proposes to extract association rules based on user interest.

MIRO consists of the following 5 steps

1. Input Ontology
2. Define User Data
3. Search Ontology
4. Generate Candidate Itemsets
5. Visualize the Results

Input Ontology: A sample dataset for electronic dataset [11] is shown in the following figure 2 which consists of 58 different items.




-
Ipod Speakers
-
Logitech s1 25i portable ipod speaker.
-
Logitech pure Fi express plus ipod dock/speaker.
-
Ipod Screen Protectors
-
Cascade screen guard mira silver apple ipod Touch.
-
Screen protector bodifender ipod nano 6 mira silver.
-
Cascade screen protector mira silver scratch guard apple ipad2.
-
MP3 & MP4 Players
-
MP3 players
-
Philips 2GB GoGear mix mp3 player.
-
Sonilex mp3 player.
-
Philips 4GB RAGA GoGear screen mp3 Player.
-
MP4 players
-
Transcends digital music player MP870 4GB.
-
7 Sunvisor TFT LCD screen monitor with MP4 DVD player.
-
LCD, LED & Plasma TVS
-
LCD TVS
-
Sony 40
-
Sony Bravia
-
Sony Bravia 32
-
Sony Bravia 40
-
LED TVS
-
New Sansung UA-40D 5003 40.
-
Sony BraviaKOL-40EX520 40.

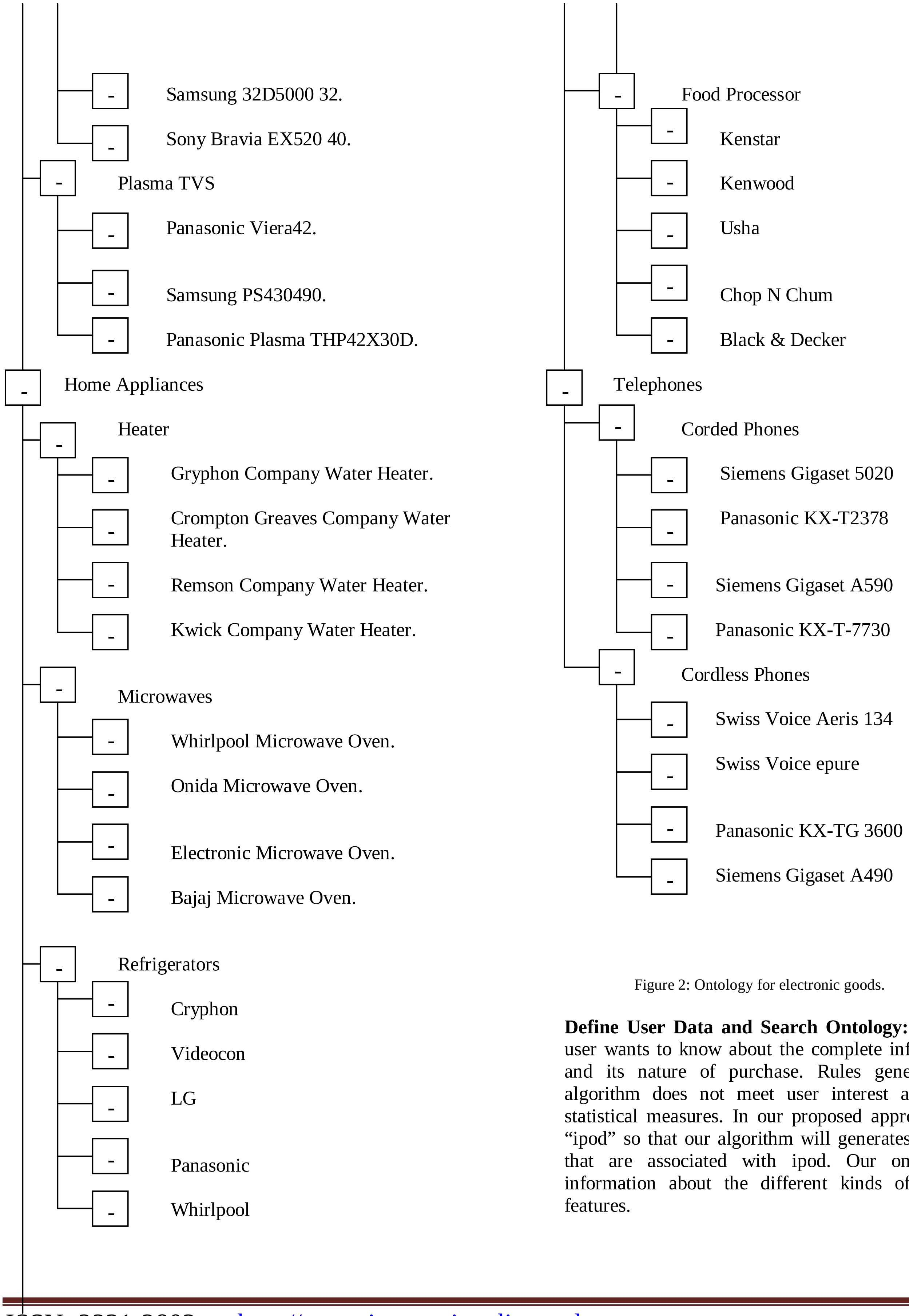
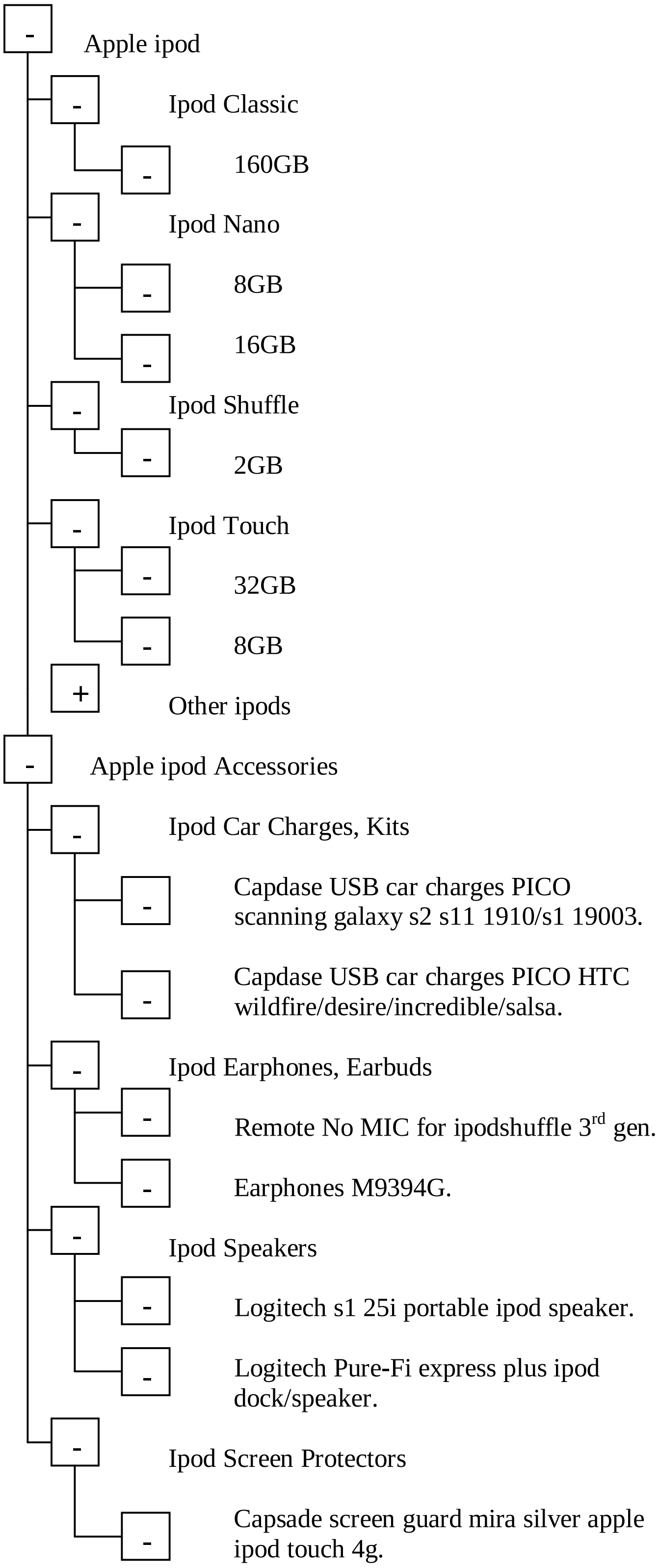


Figure 2: Ontology for electronic goods.

Define User Data and Search Ontology: For example end user wants to know about the complete information of ipods and its nature of purchase. Rules generated by Apriori algorithm does not meet user interest as it works using statistical measures. In our proposed approach user input is “ipod” so that our algorithm will generates all possible rules that are associated with ipod. Our ontology will give information about the different kinds of ipods and their features.



- Apple ipod
  - Ipod Classic
    - 160GB
  - Ipod Nano
    - 8GB
    - 16GB
  - Ipod Shuffle
    - 2GB
  - Ipod Touch
    - 32GB
    - 8GB
  - Other ipods
- Apple ipod Accessories
  - Ipod Car Charges, Kits
    - Capdase USB car charges PICO scanning galaxy s2 s11 1910/s1 19003.
    - Capdase USB car charges PICO HTC wildfire/desire/incredible/salsa.
  - Ipod Earphones, Earbuds
    - Remote No MIC for ipodshuffle 3rd gen.
    - Earphones M9394G.
  - Ipod Speakers
    - Logitech s1 25i portable ipod speaker.
    - Logitech Pure-Fi express plus ipod dock/speaker.
  - Ipod Screen Protectors
    - Capsade screen guard mira silver apple ipod touch 4g.

Table1: Initial Itemset

Generate Candidate Itemsets: Initial itemset given to our algorithm will be shown in table1.

Visualize the Results: Results are shown in the form of rules. This step can be evaluated by using data visualization technique such as bar chats, live charts etc.

IV. CONCLUSION

In order to improve the post-processing of association rules a new framework is proposed in order to generate association mining rules of user interest.

V. FUTURE WORK

Our framework should be evaluated through a scenario based analysis in comparison with other existing scenarios and a prototype based performance evaluation in terms of query response time, the precision and recall ratio, and system scalability. Our future work also concentrates on reducing the number of rules either by using filters or by using semantics.

REFERENCES

[1] A. Bellandi, B. Furletti, V. Grossi, and A. Romei, “Ontology- Driven Association Rule Extraction: A Case Study,” *Proc. Workshop Context and Ontologies: Representation and Reasoning*, pp. 1-10, 2007.

[2] A.Razia Sulthana B.Murugeswari, “ ARIPSO : Association Rule Interactive Postmining Using Schemas And Ontologies” , PROCEEDINGS OF ICETECT 2011 IEEE.

[3] [Anyanwu and Sheth, 2003] Anyanwu, K. and Sheth, A. (2003). ρ-Queries: Enabling Querying for Semantic Associations on the Semantic Web. In *WWW ’03: Proceedings of the 12th international conference on World Wide Web*, pages 690–699, New York, NY, USA. ACM Press.

[4]. Berka, P., Bruha, I.: Discretization and grouping: Preprocessing steps for data mining. In: Zytkow, J.M. (ed.) PKDD 1998. LNCS, vol. 1510, pp. 239–245. Springer, Heidelberg (1998)

[5]. Blanchard J, Guillet F, Gras R, Briand H (2005) Using information-theoretic measures to assess association rule interestingness. In: Proceeding of the 2005 international conference on data mining (ICDM’05), Houston, TX, pp 66–73

[6] [Bloehdorn and Sure, 2007] Bloehdorn, S. and Sure, Y. (2007). Kernel Methods for Mining Instance Data in Ontologies. In Aberer, K., Choi, K.-S., and Noy, N., editors, *Proceedings of the 6th International Semantic Web Conference (ISWC 2007)*, Lecture Notes in Computer Science. Springer. to appear.

[7] D.Narmadha1, G.NaveenSundar2, S.Geetha3, “An Efficient Approach to Prune Mined Association Rules in Large Databases”, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 1, January 2011.

[8] [Getoor and Licamele, 2005] Getoor, L. and Licamele, L. (2005). Link Mining for the Semantic Web, Position Statement. In *Proceedings of the Dagstuhl Seminar in Machine Learning for the Semantic Web*.

[9]. Gionis A,Mannila H, Mielikäinen T, Tsaparas P (2006) Assessing data mining results via swap randomization. In: Proceeding of the 2006 ACM SIGKDD international conference on knowledge discovery in databases (KDD’06), Philadelphia, PA, pp 167–176

[10] Hetal Thakkar, Barzan Mozafari, Carlo Zaniolo. Continuous Post-Mining of Association Rules in a Data Stream Management System. Chapter VII in Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction, Yanchang Zhao; Chengqi Zhang; and Longbing Cao (eds.), ISBN: 978-1-60566-404-0.

[11]. http://electronics.shop.ebay.in/

[12] [Jiang and Tan, 2006] Jiang, T. and Tan, A.-H. (2006). Mining RDF Metadata for Generalized Association Rules: Knowledge Discovery in the Semantic Web Era. In *WWW ’06: Proceedings of the 15th international*

conference on World Wide Web, pages 951–952, New York, NY, USA. ACM Press.

[13]. R. Natarajan and B. Shekar, “A Relatedness-Based Data-Driven Approach to Determination of Interestingness of Association Rules,” Proc. 2005 ACM Symp. Applied Computing (SAC), pp. 551-
552, 2005.

[14]. Srikant, R., Agrawal, R.: Mining generalized association rules. In: VLDB 1995: Proceedings of the 21th International Conference on Very Large Data Bases, pp. 407–419. Morgan Kaufmann Publishers Inc., San Francisco (1995)

[15] Wenxiang Dou, Jinglu Hu, Gengfeng Wu, “Interesting Rules Mining with Deductive Method”, ICROS-SICE International Joint Conference 2009.

[16]. Xin D, Shen X, Mei Q, Han J (2006) Discovering interesting patterns through user’s interactive feedback. In: Proceeding of the 2006 ACM SIGKDD international conference on knowledge discovery in databases (KDD’06), Philadelphia, PA, pp 773–778

[17] Xuan-Hiep Huynh, Fabrice Guillet and Henri Briand, “Extracting representative measures for the post-processing of association rules”, 2006 .