Application of Data Mining Techniques for Prediction and Visualization in E-commerce

Jyothi T[1], Ambika V[2]
[1,2]Assistant Professor,
GSSS Institute of engineering and technology for Women, Mysuru
Jyothit@gsss.edu.in, ambikav@gsss.edu.in

ABSTRACT.

Data Mining is the process of discovering interesting patterns from large datasets. Prediction of the current shopping trends can be done using frequent pattern mining technique. The algorithmssuch as Aproiri and FP-tree has no visualization of the frequent patterns and also consumes more time for analyzing large datasets. Works reported until now have reduced the time consumption for analyzation but does not visualize the generated patterns. This paper presents the visual exploration of frequent patterns for prediction of shopping trends using Boyer-Moore and Aproiri algorithm. Thegenerated frequent patterns are visualized into different graphs- Month Wise,Support vs Patterns, Confidence vs Patterns and Profit vs Patterns.

Keywords : Frequent Pattern, Shopping Trends, Data visualization, support and Confidence.

I. INTRODUCTION

Large data is generated in real life applications such as electronic store which are difficult to analyze using existing data discovery techniques such as Apriority or FP-Tree.One of the important area in data mining is concerned with the discovery of frequent patterns and interesting association rules. While considering the transactional databases, the patterns are extracted over a certain period of time like and at the end of the day. But, as the trends are continuously changing, the patterns extracted by previous day transactions may not suit to the present trends. Hence there is a need in extracting frequent patterns in dynamic datasets.

The standard approach to update dynamic dataset is, applying the data mining algorithms continuously for every update in the dataset. Discovering interesting patterns such as: “which items are purchased together”, from this large set of textual rules is a tedious job therefore Association Rule is used for discovering interesting relations between frequent patterns in large datasets. For example, the rule {Bread, Jam} => {Cheese} found in the sales data of a supermarket would indicate that if a customer buys Bread and Jam together, he or she is likely to also buy Cheese. Such information can be used for making decisions about marketing activities like promotional pricing or product placements. By applying the support and confidence the frequent patterns are obtained and presented in graphical form which allows visual exploration of frequent patters that predicts the efficient shopping trends by promoting visualized business analytics.

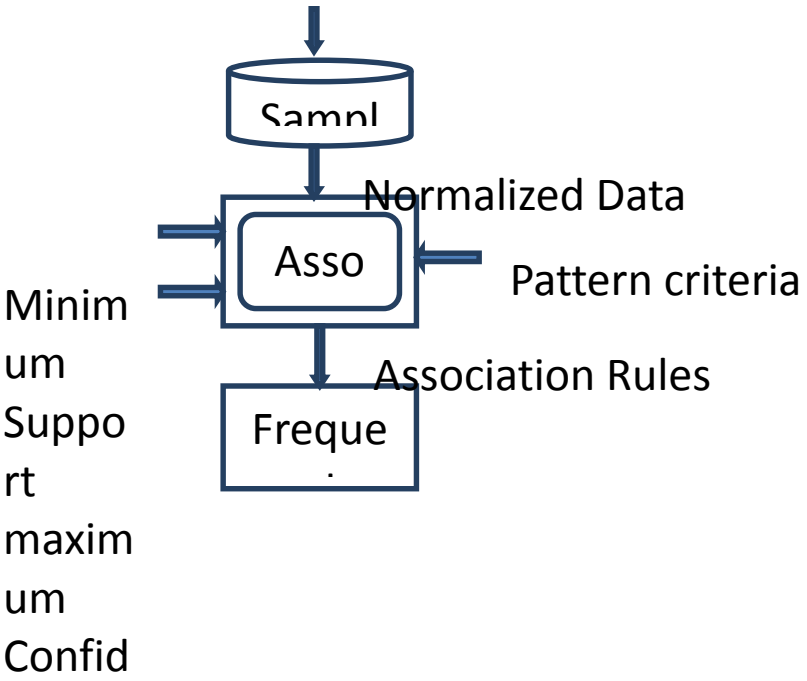

Fig1-Traditional approach for association rule mining

The problem with traditional approach for association rule mining of patterns is difficult, as algorithms are very time and memory consuming. And the number of patterns extracted by current solutions is too large to be easily handled by end users.Fig1 presents the traditional approach to association rule mining where support and confidence are used to produce best rules.Proposed visualization technique uses Line graphs which are useful in displaying data or information that changes continuously over time and the Bar Charts for visual impression of the distribution of data. Bar charts can display large amounts of data that are difficult to understand in a tabular or spreadsheet form. Therefore line graphs and Bar charts are more suitable for visualizing the generated frequent patterns in dynamic datasets.

II. RELATED WORK

Association rules and frequent patterns can be discovered by using 3 types of measures such as i) Objective Measures (ii) Subjective Measures and (iii) Semantic Measures. Objective measure is a data-driven approach for evaluating the quality of association patterns. Subjective measures generally operate by comparing the beliefs of a user against the patterns discovered by the data mining algorithm. SSFPOA extracts and clusters semantically similar frequent patterns. It uses both domain dependent and domain independent ontologism, and considers the entire path to conform the semantic similarity between elements along with their structural information [1]. Visualize that uses a radial layout to show frequent patterns. A legible right-side-up graph to one user may become an illegible upside-down graph towards another user [2].

FpViz represents each frequent pattern visualized in graph where horizontal line represents k number of items and vertical line represents k number of nodes where each node represents an item within the frequent pattern [3]. FIsViz visualize represents frequent pattern by a plotline. Domain items are taken along one axis and frequency is taken on the other axis. Problem with FIs Via is that it does not scale well. The location of the polyline indicates the exact frequency of the pattern explicitly [4]. CloseViz proposes a visualize –in order to show only closed frequent patterns. Existing visualization systems were not designed to show frequent patterns and particularly semantically similar frequent patterns [5].

III. PROBLEM STATEMENT

The traditional approach generates inefficient patterns which are not useful for shopping trends prediction leading to error prone results. The major drawback of traditional approach is not visualizing the generated patterns. Whereas the proposed system generates the efficient frequent patterns and predicts the shopping trends by applying association rules to dynamic datasets by visualizing the generated frequent patterns using different graphs to obtain shopping trends in E-commerce.

IV. PROPOSED SYSTEM

The proposed system provides the visual exploration of frequent patterns for prediction of shopping trends using Boyer-Moore and Apriority algorithm. The generated frequent patterns are visualized into different graphs- Month Wise, Support vs. Patterns, Confidence vs. Patterns and Profit vs. Patterns.

For visualization of the generated frequent patterns the proposed system uses line graphs which are useful in displaying data or information that changes continuously over time and the Bar Charts for visual impression of the distribution of data. Bar charts can display large amounts of data that are difficult to understand in a tabular or spreadsheet form. Therefore line graphs and Bar charts are more suitable for visualizing the generated frequent patterns in dynamic datasets.

V. METHODOLOGY

Algorithm-Apriority and Boyer-Moore Algorithm

Step 1: Scanning of Data- Scan the dataset (storage servers such as database or data stored in cloud, excel sheet, files).

Step 2: Data Cleaning- removing the irrelevant information and retaining relevant information for mining.

Step 3: Selection of Data (extract the target data for mining)- Extraction of previous customer transactions data set .

Step 4: Preprocessing of Data- extracting the customer previous transactions information based on the customer age, month wise, season wise, customer occupation wise etc.

Step 5: Calculating the support for distinct items [C1].
Step 6: Generate L1 (Frequent one item set).
Step 7: Use Lk-1, join Lk-1 to generate the set of candidate k - item set.
Step 8: Scan the candidate k item set and generate the support of each candidate k – item set.
Step 9: Add to frequent item set, until C=Null Set.
Step 10: For each item in the frequent item set generate all non empty subsets.
Step 11: For each non empty subset determine the confidence. The output of Apriority algorithm is given as input to the Boyer-Moore algorithm. If confidence is greater than or equal to this specified confidence – efficient pattern discovery
Step 12: Calculate the profit level for each pattern and extract patterns based on descending order.

A. Terminologies used in Algorithm

ITEM SET: The set of an item which is being sold.
SUPPORT: The support of the item set. It is the relationship between the total number of transaction containing the item with the total number of transaction in the data set.

Support(A->B)=The total number of transaction containing A and B / The total number of transaction in the dataset.
CONFIDENCE: Confidence is defined as a relationship between the total number of transaction containing the frequently bought items with the total number of transaction Containing only A.

Confidence(A->B) = The total number of transaction containing A and B / The total number of transaction containing only A.

VI. DATASETS

The datasets are storedin SQL server database in the form of tables containing Admin details, Customer details, Item category, Item subcategory, Item details, Customer transactions, Transaction details, Problem Details, Strong rules, Feedbacks.

VII. EXPERIMENTAL RESULTS AND DISCUSSION

The frequent patterns are extracted by applying the association rules like the support and confidence, interesting frequent patterns generated from the dynamic datasets are then visualized in the form of graphs using Boyer-Moore and Aproiri. Visual exploration of frequent patters predicts the shopping trends by promoting visualized business analytics. Fig1 shows the month-Wise visualization, Fig2 shows the support vs the patterns, Fig3 shows the visualization of confidence vs the frequent patterns and Fig4 shows the profit-wise visualization of shopping trends.

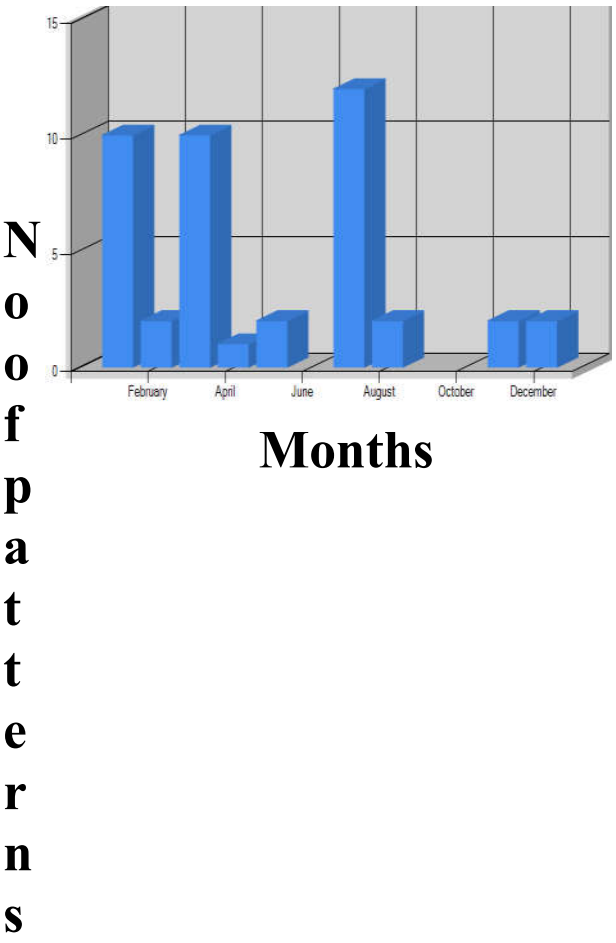

Fig1- Month wise visualization of shopping trends using column graph.

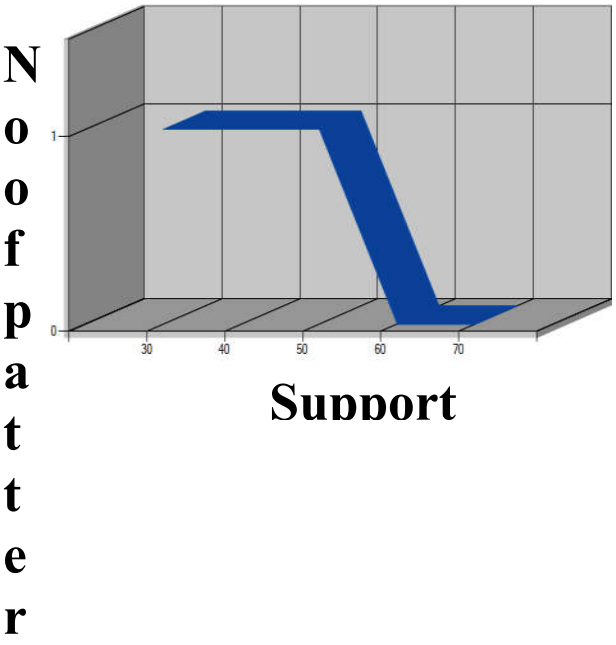

Fig2- Line graph visualization of shopping trends based on support.

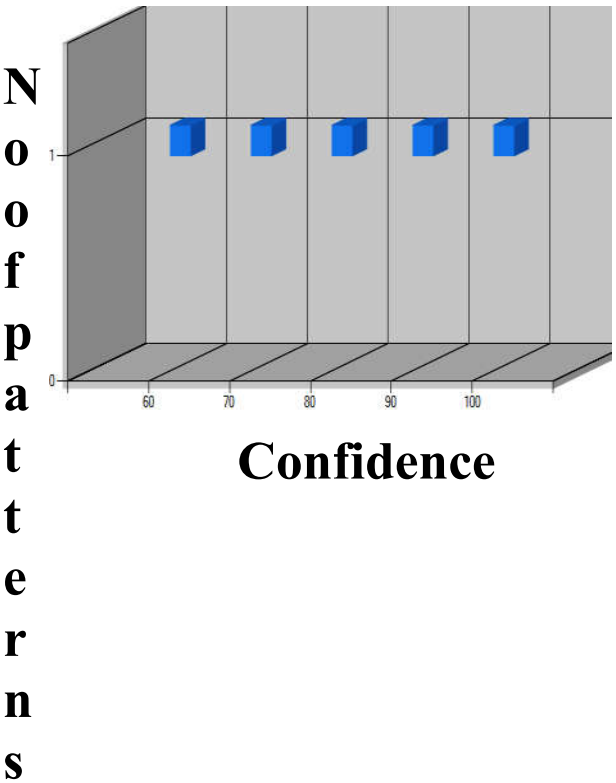

Fig3- point graph visualization of shopping trends based on confidence.

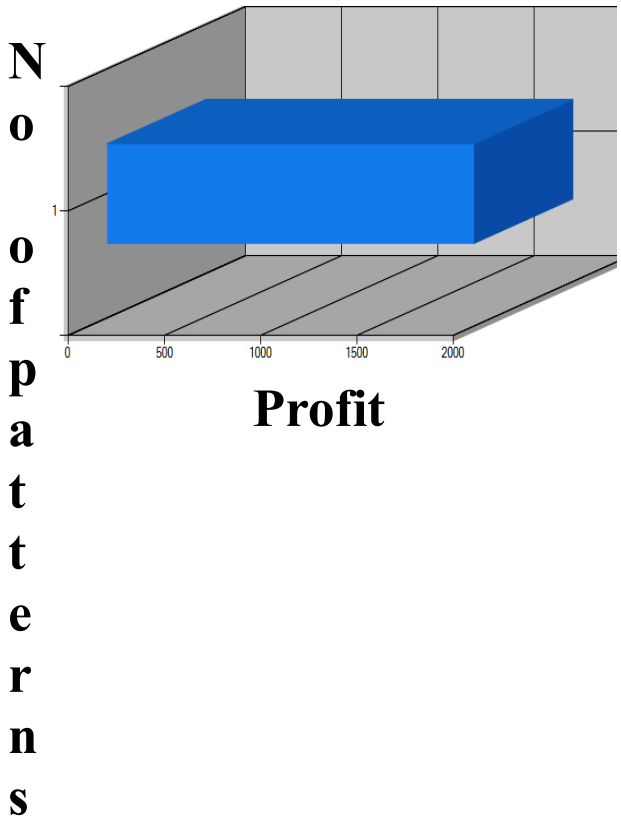

Fig4- Profit wise visualization of shopping trend using bar graph.

CONCLUSION

Visualization helps business analytics to easily compare importance of each item-set. The interactive features such as support and confidence values, considering the transactions for a given period of time, colors for display would help in different kinds of patterns to be extracted. According to the rapidly changing present shopping trends, it is essential to consider dynamic Datasets for extracting frequent patterns. Existing data mining algorithms are very time consuming and they generate very huge number of frequent patterns, hence we extend SSFPOA algorithm which results the semantically similar frequent patterns at higher levels of abstraction.

The predicted and visualized shopping trends are used for different analysis which provides the information of product trends based on different months using the month-wise graph. The support graph, confidence graph and the profit graph provides the product combinations purchased that helps businesses to promote their most profitable products to maximize the business profit. The future work concentrates on providing visualizations of patterns based on season-wise and the comparison of performance of various visualization techniques.

REFERENCES

1]. S. Vasavi, S. Jayaprada, V. Srinivasa Rao, "Extracting Semantically Similar Frequent Patterns Using Ontologies", *Proceedings of the Second international conference on Swarm, Evolutionary, and Memetic Computing* - 2011.
2]. Carson Kai-Sang Leung and Fan Jiang. "RadialViz: An orientation- free frequent pattern visualizer*", In Proceedings of the 16th Pacic-AsiaConference on Knowledge Discovery and Data Mining.* Springer, 2012.
3]. Carson Kai-Sang Leung, Christopher L. Carmichael, "FpViz: A Visualizer for Frequent Pattern Mining", ACM 978-1-60558-670-0/09/06, VAKD '09
4]. C.K.-S. Leung, P.P. Irani, and C.L. Carmichael. "FIsViz: a frequent itemset visualizer". In Proc. PAKDD 2008.

5]. Christopher L. Carmichael, Carson Kai-Sang Leung, "CloseViz: Visualizing Useful Patterns" UP' 10, 2010, ACM 978-1-4503-0216-6/10/07.
6]. Jing He,"Advances in Data Mining: History and Future", *Third international Symposium on Information Technology Application*, 978-0-7695-3859-4/09 IEEE 2009 DOI 10.1109/IITA.2009.204.
7]. M.S. Chen, J. Han, and P.S. Yu."Data mining: An overview from database perspective", *IEEE Transactions on Knowledge and Data Eng*., 8(6):866-883, December 1999
8]. Venkatadari M., Dr. Lokanataha C. Reddy, "A Review on Data Mining From Past to Future", *International Journal of Computer Applications,* pp.19-22, vol. 15, No. 7, Feb 2011
9]. C.C. Aggarval, P.S. Yu "Finding localized associations in market basket data ",*IEEE Transactions on Knowledge and Data Engineering*, 14 (2002), pp. 51–62
10]. S.R. Ahmed ,"Applications of data mining in retail business"*Information Technology: Coding and Computing*, 2 (2004), pp. 455–459