International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August 2019

Cluster Centric Technique for Association Rule Mining(C-ARM) For High Precision and Coverage in Web Page Recommendation

Manikandan R, Saravanan V

Abstract: ***Association Rule Mining (ARM) is known for its popularity and efficiency in the data mining domain. Over the recent years, the amount of data that gets accumulated in the internet is getting increased exponentially over time. The data available so are stored in online and are retrieved when a user requests for the same through key words with the help of a search engine. The important task of the search engines are to present the appropriate web pages that an user is expecting and in the modern times, The need of the hour is to recommend web pages to the users that he is interested in. This made the web page recommendation an important and vital task. Although many of the researchers are in the preliminary task of developing such systems, we in this research propose a recommendation model in which different users are interested upon a common item or domain by using the ARM concept. The data patterns that are in common are identified using the ARM and further these are clustered on a form of hierarchy .The clusters makes the recommendation system to easily identify the user group and based on the group, the pages are recommended, The experimental analysis are discussed and found to be efficient than the available methods in terms of computation time and reliability.***

Keywords: Association Rule Mining (ARM) , Clustering, Data Mining, Web page recommendation , and User domains.

I. INTRODUCTION

I. Introduction

The drastic growth of the unstructured data in the internet is becoming out of control of the users and data makers. The ability of a human being to interpret the data and the thirst for information in the internet is also fast growing. The task of providing the needed information for the web users has been a trivial work for the serviced providers. The dynamic change in the needs of a user is also a major factor to be concerned for predicting the needs of the user. As most of the web data are unstructured, primitive techniques fail to retrieve useful information and hence the need to analyze the user behavior and browsing pattern has been essential to provide web recommendations that are personalized for a particular user.

Revised Manuscript Received on August 30, 2019.

* Correspondence Author

Manikandan R* , Research Scholar , Anna University, Chennai. Email : mani4gift@gmail.com

Saravanan V, Professor and Dean, Department of Computer Applications,Sri venkateswara College of Computer Applications and Management, Coimbatore.

The need of the hour was to provide a tailored browsing experience which is capable to handle the problem of overloaded data in the web. Applying data mining techniques in the usage data has become a popular way to understand the behavior of an user and termed as web usage mining[1]. Traditional methods are limited due to the absence of the knowledge of the Domain concerned as the focus was in usage mining alone. A a result the browsing patterns evolved were of poor quality. The patterns evolved fails to understand and predict the users' insight and interests and leading to poor recommendations.

Association rule learning is a widely used technique to discover interesting relations in between a large set of variables particularly when the size of database is large. Rules that are strong are identified with the aid of various measures of interests [1]. Rakesh Agrawal et al.[2] used the concept of rules that was introduced and put to practice in finding the regularities in purchasing pattern of users in a large transactional data which was captured by a POS system.By taking the original theory of association proposed by Agrawal et al.[2] a definitive function for the ARM is defined as follows:

- Let there be n attributed of binary type in a set S termed as items.
- Let D= $\{t1,2,\ldots tm\}$ is the transaction set of a database. .
- Assumed that every transaction has a unique ID and has a subset of the items in I.
- A rule is considered as an implication and the item set X and Y are termed as antecedent (LHS) and consequent (RHS) of the rule.

.The rules that are associative are generally expected to satisfy values prescribed by the users and are termed as minimum-support and minimum confidence at any given point of time. The steps involved in generation if a association rule are given as:

- All the item sets that are more frequent in the given database are identified by applying a user defined minimum support
- These sets of item and the other user specified confidence value are made use of to form rules.

The focus is brought on the second step as the former is simple and understood. Although many algorithms have been proposed for associative rule generation such as Apriori, Eclat and FP growth, trhese generally do only the 50% of the required task as it forms only the frequent sets of item.

The major work lies in the second step or phase where the intention is to derive at rules from the frequent itemsets.

The works carried out in the research paper is as follows

a) A detailed study of the recent issues and challenges in a recommendation model.

b) The time factor for searching is identified as one of the major issues and the methods to overcome the same has been studied by making a survey on the available ARM and Clustering techniques.

c) Proposed a novel model that reduces the complexity of time by using the clustering technique

d) Have proved that the proposed method are more effective when compared to the existing methods in terms of complexity involved in response time.

The paper is structured as follows: Section II depicts the previous works related to ARM, section III gives a lucid explanation of the proposed methodology and the results of the experiment carried out are discussed in Section IV and Section V gives the conclusion and future works.

II. a) Related works on ARM

The following section provides a gist of various studies on ARM by different researchers.

The method in [3] is based on the concept of pattern mining and are use in the market basket analysis to generate rules for the items that may be brought together . the main objective being the improvisation of the quality of the database and there by achieving a great deal of decisions support queries.

The methodology in [4] proposed apriori algorithm for getting more candidate sets. This suffers from the drawback of having to be run several passes over the complete database and also ends up in finding them non frequent. The author in [5] suggests for a association rule mining which is in many dimension for identifying the correlation in between the various attributes. The major advantage being that all the data types are considered and the disadvantage being the huge amount of information overload which makes the efficiency very poor and also tend to have a large time complexity. The FUP (Fast UPdate) method [6] was introduced to deal with new item problem ir when a new transaction is recorded. The problem with this method is the inability to update when deletion and updation is done on a same transaction.

The author in [7] proposed a Association rule mining based on various levels. This focuses on the mining of rules that are strong and which are among the various levels of abstraction. For instance the rules can be made of milk and jam on one level and on the other level association can be made of drink and fruits. Various constraints are applied to the available methods so as to make it efficient by presenting only the rules that are of users interest instead of all [8].

II. b) Related works on reducing time complexity

The author in [9] removed the process of generating candidate and thus making only two passes over the complete DB which resulted in generation of frequent item sets with less time complexity. In [10], RARM proved to be faster than FP-Tree with the end results. The SOTrieIT is proved to generate both large and quick item-sets by not making the database scanned for the second time. A novel approach termed as Inverted Hashing and Pruning (IHP) [11] was introduced for mining rules in a transactional database. The results based on the performance metrics proved that the IHP outperforms the existing methods particularly when long transactions are involved. Fuzzy grids based rules mining algorithm (FGBRMA) was proposed in [12] for the creation of association rules that are fuzzy from a RDBMS. This method comprised of two stages namely

1. For generation of fuzzy grids
2. To generate fuzzy rules

A theoretical proof was presented over its effectiveness in a particular database.

II c) Related works on Clustering

The author in [13] have proposed a method based on clustering called the HBM (Hierarchical Bisecting Medoids Algorithm) for clustering that are based on the users' session of navigation. These navigation are then grouped and clustered to identify association rules through which similar type of students are predicted in the future for e-learning of a course and to recommend them with the same. The efficiency and the effectiveness are compared with that of the existing method of clustering.

The author in [14] proposed a method for building a classifier that was based on the association rule mining in an extended scenario. A threefold method is proposed. First to apply the gain in information, second to integrate the process and third to bring in the rules to avoid redundant rules and conflicts. Classification Association Rule Mining (CARM) [15] obtains the necessary rules from the training set of data that are classified previously. This makes that the rules that are generated will be decided by the parameters of the ARM that is deployed by the algorithm itself. This outperforms its predecessor[CARM] in terms of accuracy. The weighted association rules (WARs) [16] are brought in mainly because of that the item's importance are different. NARs are the vital factor in decision support systems. But the misleading rules occur and some rules are uninteresting when discovering positive and negative weighted association rules (PNWARs) simultaneously. A fuzzy system is proposed in [17] to address the limitation. Here the association which are both direct and indirect are considered as it has many minimum supports and confidences that helps greatly to resolve the limitation.

The conventional algorithms that are used for the data mining of ARS are normally built upon attributes database that are binary in nature, with couple of imitations [18]. Firstly, it cannot concern quantitative attributes; secondly, it treats each item with the same significance although different item may have different significance. In [19] A Associative Rule Mining problem that is theoretical is addressed with the aid of multi-objective prospect by presenting an MOPAR algorithm to optimize the rules that are generated. These discovers the ARs in a numerical fashion. More efficient rules are identified by taking more objectives. This method proved to be more efficient in terms of many parameters that includes confidence and comprehensibility. Finally, the Pareto concept for optimality is used to obtain , the best ARs .

II. PROPOSED TECHNIQUE

The following segment explains the proposed methodology for achieving a better recommendation. The logs from the web servers are an important source from where useful patters can be identified of a users behavior. This makes them to navigate to more appropriate pages in a definite server.

ARM is mostly used in case of web transactional process that depicts the relation between the pages and interests based on the factor called hit ratio. Below mentioned are the steps followed in the proposed method.

a) Input: The Weblogs that are selected from the server.
b) The collected data is subjected for preprocessing initially.
c) Redundancy is avoided by performing data cleaning.
d) Selection of appropriate data with relevant features is carried out for constructing the recommendation model.
e) These are stored in a database online
f) Clustering is carried out using a hierarchical manner for data segmentation and further assigns a index for classifying further.
g) The data that are clustered are made use for the estimation of minimum support and confidence.

The ARMs generated will be of similar patterns and the recommended pages are evaluated.

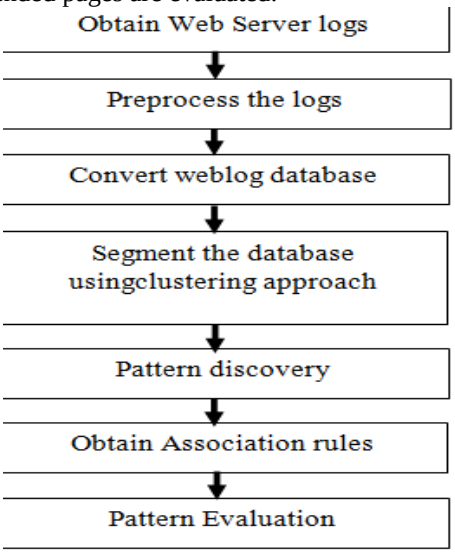

Figure 1 – Steps in the proposed methodology

III. EXPERIMENTAL SETUP AND ANALYSIS

The following section describes the experimental setup and analysis of the results.

a) **Web server Logs:** These are collected from a web server and of the type .log

b) Preprocessing: The incomplete and redundant requests are removed in order to achieve better accuracy.

c) **Creation of Database:** The logs can not be used directly for processing ARMs in the experiment. Hence it is converted to MySql data base.

.d) **partitioning the database**: The hierarchical clustering is done based on the count of the support. The users interest are found using this,.

e) **Discovery of patterns**: Similar patterns that are present inside a cluster are identified.

f) **Association rules:** It describes the relationship between different itemsets.

g) **Evaluating the patterns**: It is done so as to interpret the model for recommendation.

Since, Association rule mining is the basic concept of the proposed web recommendation model. It is measured from support and confidence values of the predicted rules.

a) Support: The support of the web page (P) is calculated by the transaction's proportion in which a relevant web page is present.

b) Confidence: The confidence of the rule is defined from the eqn (1):

$$Conf\ Pi{-\!\!-\!\!\rightarrow}Pn = (Pi \cup Pn) supp(Pi) \quad (1)$$

The metrics by which the proposed method is evaluate are the

a) Precision

It is the prediction of information that is accurate for recommendation for all the test users.. It is given as in eqn (2):

$$Precision = (T\ p \cap\ )(p) \quad (2)$$

Where R(p) is the recommendation set and T(p) is the users session. It is subjected to change based on the recommended pages.

b) Coverage

Coverage is the proportion of relevant recommendations to the all pages that should be recommended. It given as in eqn(3)

$$Coverage\ Pn \rightarrow Pi = Support\ (Pi) \quad (3)$$

The table and the figures visually shows the performance of the proposed system when compared with the existing methods. The Comparative analysis of the rate of precision among the proposed and the existing method are categorized based on the number of pages that are ranked.

Tables 1 and **2** shows the readings we got during our practical analysis and **Figures 2** and **3** shows the graphical view of the results.

No of Pages	Precision	
	Proposed	Previous
1	100	100
2	50	33
3	100	100
4	100	87
5	80	60
6	58	50
7	57	50
8	85	62
9	33	27

Table 1 : Precision comparison readings

No of Pages	Precision	
	Proposed	Previous
1	50	50
2	66	66
3	65	16
4	100	50
5	100	83
6	57	50
7	100	83
8	62.5	16
9	33	20
13	62.5	39

Table 2 : Coverage comparison readings

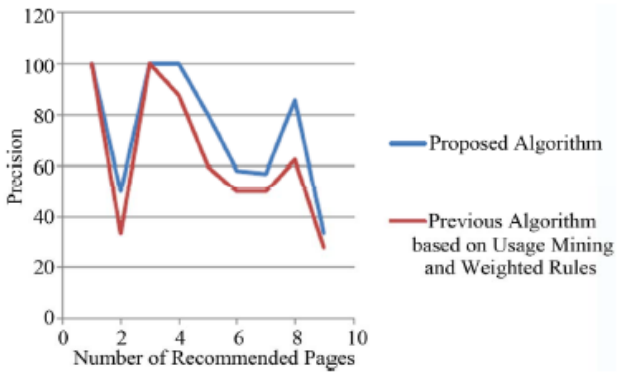

Figure 2: Comparative analysis of Precision

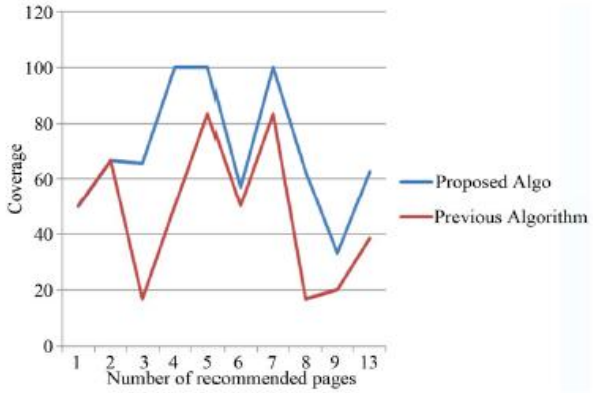

Figure 3 : Coverage comparative analysis

IV. CONCLUSION

Web mining is opting out as an emerging and most researched topic in the recent past. The need of more appropriate pages which are matching the user interest is high. Although the data getting accumulated in the internet gets doubled every two year, constant researches are still going on to make a efficient recommendation model. The proposed ,metod makes uses of the ARs and the clustering concept to develop a more efficient recommendation model which is apt for the user. These make use of the associative rules generated and then forming a cluster. The experimental results have proved that the performance of the proposed system is better than the previous techniques used in terms of the precision and coverage. The future work aims to extend this approach on a distributed environment where the web pages are captured from different domains and servers but having a common functionality of the users request.

REFERENCS

1. William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus, Knowledge Discovery in Databases: An Overview, *AI Magazine,13(3)* ,1992.
2. Jochen Hipp, Ulrich Guntzer, and Gholamreza Nakhaeizadeh,Algorithms for Association Rule Mining – A General Survey and Comparison, *Volume 2, Issue 1* ,2000,pp58 .
3. Agrawal.R, Imielinski.T, and Swami.A.N, Mining Association Rules between Sets of Items in Large databases, *ACM New York, NY, USA*, 1993, 207-216.
4. R.Agrawal, and R.Srikant, Fast algorithms for mining association rules, 1994, 487-499.
5. R.Srikant, and R.Agrawal, Mining quantitative association rules in large relational tables, *ACM Press*, 1996, 1-12.
6. D. W.Cheung, S. D.Lee, and B .Kao, A general incremental technique for maintaining discovered association rules. In Database Systems for Advanced Applications, 1997, 185-194.
7. J. Han, and M.Kamber, *Data Mining Concepts and Techniques* (Morgan Kanufmann, 2000).
8. J.Pei, and J.Han, Can we push more constraints into frequent pattern mining?,*ACM Press*, 2000,350-354.
9. J.Han, and J.Pei, Mining frequent patterns by pattern-growth: methodology and implications, *ACM SIGKDD Explorations Newsletter 2, 2*, 2000, 14-20.
10. A.Das, W.-K.Ng, and Y.-K.Woon, Rapid association rule mining, ACM Press, 2001, 474 - 481.
11. John D. Holt, and Soon M. Chung, M ining association rules using inverted hashing and pruning, *Elsevier, Volume 83, Issue 4*, 31 August 2002, 211-220.
12. Yi-Chung Hu, Ruey-Shun Chen, and Gwo-Hshiung Tzeng, Discovering fuzzy association rules using fuzzy partitionmethods, Elsevier, Volume 16, Issue 3, April 2003, 137-147.
13. Feng-Hsu Wang, and Hsiu-Mei Shao, Effective personalized recommendation based on time-framed navigation clustering and association mining, Elsevier ,Volume 27, Issue 3, October 2004, 365-377.
14. Guoqing Chen, Hongyan Liu, Lan Yu, Qiang Wei, and Xing Zhang, A new approach to classification based on association rule mining,Elsevier, Volume 42, Issue 2, November 2006, 674-689.
15. Frans Coenen, and Paul Leng, The effect of threshold values on association rule based classification accuracy, *Elsevier, Volume 60, Issue 2*, February 2007, 345-360.
16. He Jiang, Yuanyuan Zhao, and Xiangjun Dong, Mining Positive and Negative Weighted Association Rules from Frequent Itemsets Based on Interest, *IEEE ,vol.2*, 2008, 242,245.
17. WeiminOuyang, and Qinhua Huang, Mining direct and indirect fuzzy association rules with multiple minimum supports in large transaction databases, *IEEE, vol.2*, 2011,947-951.
18. Anjana Gosain, and Maneela Bhugra, A Comprehensive Survey of Association Rules On Quantitative Data In Data Mining, *IEEE*, 2013.
19. Yiyong Xiao, Yun Tian, and Qiuhong Zhao, Optimizing frequent time-window selection for association rules mining in a temporal database using a variable neighbourhood search, *Elsevier Volume 52*, December 2014, 241-250.
20. Vahid Beiranvand, Mohamad Mobasher-Kashani, and Azuraliza Abu Bakar, Multi-objective PSO algorithm for mining numerical association rules without a priori discretization, *Elsevier, Volume 41, Issue 9*, July 2014, 4259-4273.

AUTHORS PROFILE

Manikandan R , an academician with 7 years of teaching experience and in research . Area of interest includes Data mining and Big Data.

Dr. V Saravanan has a PhD degree in the computer science and has over 15 years of academic experience .His research area includes Data Mining, Artificail Intelligence, data cleaning and Software Agents.