International Journal of Scientific Research in Computer Science and Engineering
Vol.5, Issue.6, pp.13-18, December (2017)

Survey Paper

E-ISSN: 2320-7639

Fuzzy Association Rule Mining- A Survey

Pramod Pardeshi[1*] and Ujwala Patil[2]

[1*] Dept. of Computer Engineering, R. C. Patel Institute of Technology Shirpur, MS, India
[2] Dept. of Computer Engineering, R. C. Patel Institute of Technology Shirpur, MS, India

**Corresponding Author: pramod.rcpit@gmail.com*

Available online at: www.isroset.org

Received: 12/Nov/2017, Revised: 20/Nov/2017, Accepted: 17/Dec/2017, Published: 31/Dec/2017

***Abstract*—** The World Wide Web has become a huge repository of the hypertexts and documents. The rapid growth of web logs and texts has gained a lot of attention from the researchers for extracting the interesting rule for designing of the web pages, drawing the customer preference, analysing the customer behaviour and decision making for serving the organizations with better services. Such decisions are made by analysing different web parameters such as the server log, registration information, access time, session period, page hits and other relative information left by user. This paper presents a survey on various techniques such as fuzzy logic and rule mining for finding the customer behaviour that helps in better decision making and enhancing the performance of the system.

***Keywords*—** Video watermarking, feature selection, rough set theory, motion vectors, particle swarm optimization.

I. INTRODUCTION

With The advancement of the Internet has attained the peak of information gathering and exchange among a wide variety of users working in different domains. With significant increases in the amount of the data and their users, the World Wide Web has evolved as a network of data comprising of a variety of application texts and documents. The huge amount of data available over a number of websites has increased in level of complexity in managing and searching data for the users. Web mining has gained a lot of attention in the area of data mining because of its valuable contribution towards the end users and has developed new ideas for efficient accessibility of the data in less amount of time. The data mining is Off-Line process having its data stored in the data warehouses whereas the web mining is On-Line process where the data is stored in the server database and web logs [1]. The large amount of data generated in log files which is very significant because many times users repeatedly access the same type of web pages and the record is maintained in log files. These series can be taken for as a web access pattern which is helpful to find the user behavior. Through this behavior information, we can find out the correct user next request forecast that can reduce the browsing time of web pages.

A. Web Mining

Web mining has three distinct categories, namely [2], Web content mining, Web structure mining and Web usage mining as shown in Figure 1. The content data is a collection of facts that appears on a web page. It may consist of text, images, audio, video and lists or tables. The content mining deals with the extraction of the information from the contents of web documents. Secondly, a webpage also contains nodes and hyperlinks for connecting pages. Web structure mining is the process of discovering the various structure information on a web page. Another type web mining is the Web usage mining that discovers the patterns of the data, including the user log with their IP address, page reference and access time. Usage data contain the identity of web users along with their browsing behaviour.

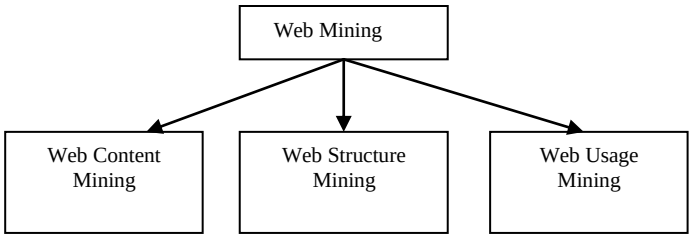

Figure 1 : Taxonomy of Web Mining

A temporal association rule relates associates the items from the same transactions. The concept of general temporal association rules has been proposed where the items are allowed to have different exhibition periods, and their supports are made in according to with their exhibition periods. A temporal association rule is find from base cubes rather than of item sets as long as it achieves the minimum support and strength.

B. Web Usage Mining

There are three basic tasks for the Web Usage mining such as preprocessing, pattern discovery and pattern analysis [3].

Pre-processing: The pre-processing is performed for the incompleteness of the data available on the web. It is a process of converting the usage, content and structured information into a data abstraction for pattern discovery. The session initiation with the user requested and the session, leaving time of the user is recorded tracking peak usage of the web. The exact content served as a result of each user is often available from the requested fields in the server logs that helps in accessing server log information content.

Pattern Discovery: The different methods are performed for mining the activities on the Web domain. Statistical techniques are applied for extracting the information about the visitors of a web page having the variables such as page views, viewing time and length of navigation path. Association rule generation is done to relate the most often visited pages in a single session server. A clustering technique tends to establish the group of users exhibiting the similar browsing patterns. Such information is helpful for analyzing the market segmentation for the E-commerce applications.

Pattern Analysis: The aim of pattern analysis is to extract the interesting rules. These rules define the behaviour of the user in a server session. These various techniques such as graph patterns or coloring of the patterns describe the various users from the different domains and their usage to the web data.

C. Association Rule Mining

Association rules are one of the main techniques of data mining [4]. It finds frequent patterns, associations, correlations, or casual structures on sets of items or objects in transactional databases, and other information repositories. The volume of data is increasing dramatically as the data produced by day-to-day activities. Therefore, mining association rules from massive amount of data in the database is curious for many industries which can help with many business decisions creating processes, such as cross-marketing, basket data analysis and promotion assortment.

Let I is a set of items. A set X = {i1, . . . , ik} $\subseteq$ I is known an item set, or a k-item set if it contains k items. A transaction over I is a couple T = (tid, I) here tid is the transaction identifier and I is an itemset. A transaction T = (tid, I) is called to support an itemset X $\subseteq$ I, if X $\subseteq$ I. A transaction database D over I is a set of transactions over I.

An itemset is called frequent if its support is no less than a provided absolute minimal support threshold min abs, with 0 ≤ min abs ≤ |D|.

An association rule is a expression of the form X $\Rightarrow$ Y , here X and Y are item sets, and X ∩Y = {}. Such a rule convey the association that if an transaction have all items in X, then that transaction also contains all items in Y. X is called the body or antecedent, and Y is known the head or consequent of the rule. The support of an association rule X $\Rightarrow$ Y in D is the support of X $\cup$ Y in D, and likewise, the frequency of the rule is the frequency of X $\cup$ Y.

The confidence or exactness of an association rule X $\Rightarrow$ Y in D is the conditional probability of having Y contained in a transaction, given that X is contained in that transaction:

$$\text{Confidence } (X \Rightarrow Y,D) := P(\Rightarrow X) = \leq \gamma \frac{\text{support}(X \cup Y, D)}{\text{support}(X, D)}$$

The rule is known as confident if P(Y |X) exceeds a given minimal confidence threshold, with 0 1.

D. Fuzzy Association

A elementary methodology based on classical association rule mining is Fuzzy Association Rules mining. Whenever data set having a certain range of values then it might possible to face the sharp boundary problem.

Suppose we have three range of marks of one examination.

F(x) is a function such that

0<x<=35 then f(x) = fail students

35<x<=80 then f(x) = average students

x>80 then f (x) = poor students

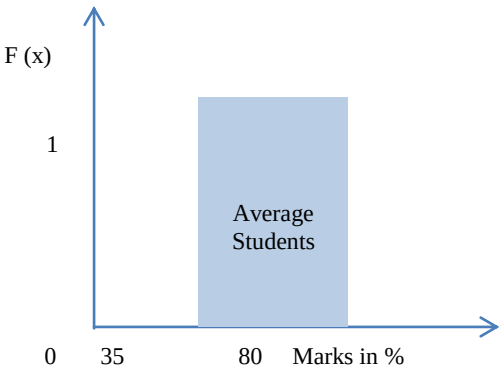

Figure 2: Example of sharp boundary problem

From the given scenario now suppose a student got 79.8% of marks then he is a average student. But if he got 80.1 Marks then the tag will good. But the student who got

79.8 % of marks is also a good student, which is not

The huge difference between both the conditions. Figure 2

display the scenario for the boundary shape problem, which happened above.

There are some basic way to solve the sharp boundary problem.

Quantitative approach

Fuzzy Taxonomic Structures

Approximate Item set Approach

To solve the sharp boundary problem by using Quantitative approach divides the variable marks into three fuzzy sets. The fuzzy sets and their members functions will have to be defined by a domain expert. For easy illustration, we will just describe the borders of the sets and split the overlying part equally between the so generated fuzzy sets. For an illustration, we will use the following borders for the fuzzy sets of the variable marks: Fail={0-35}, Average students={33-70}, good students={70-8}. The generated fuzzy sets is shown in Figure 1 . For all areas having no overlap of the sets, the support will be 1 for the actual item set. If there is an overlap, the membership can calculate by using the borders of the overlying fuzzy sets. The added support will here always sum up to 1.

II. LITERATURE REVIEW

A numerous work has been proposed for mining temporal association rules. There have been different approaches such as rule mining, clustering, fuzzy logic, etc. done for finding the usage of the web data.

In 2012 Rajni pamnani et al. [2], The basic restriction of existing Personalization systems is the loosely coupled integration of the Web personalization system with the Web server ordinary process. SUGGEST is completely online and incremental, and it is aimed at providing the users with information about the pages they may find of interest. It bases personalization on a user's classification that evolves related to the user's requests. Usage information is represented by means of an undirected graph whose nodes are associated to the identifiers of the accessed pages, and each edge is associated to a measure of the mutual relation existing between nodes (pages). This graph is incrementally changed to retain the user model update. In the model the "interest" in a page does not based on its contents, but on the order by which a page is visited during a session. Therefore, to weight each edge of the graph we introduced a novel formula:

$$Wij=Nij/max(Ni, Nj).$$

where Nij is the number of sessions including both pages i and j, Ni and Nj are the number of sessions containing only page i or j, respectively. Dividing Nij by the maximum between single existence of the two pages has the output of refined internal pages from the so called index pages. The index pages are those that do not generally contain useful content and are only used as a starting point for a browsing session. We conclude to consider index pages to be of too little interest as potential suggestions because they are very likely to be included too many sessions. The index pages are used in different works to present the results of the personalization phase. In these cases index pages are not just used to identify potentially useful information, but just to present the especially results.

In 2013 Stephen G. Matthews et al. [5], proposed a GA-based approach with an enhanced fitness function. The chromosome for the approach where each different set of parameters is composed lower temporal endpoint, upper temporal endpoint, uniform resource locator, the linguistic label for the page view of the URL, the lateral displacement of linguistic label and the antecedent with consequent are labelled. The support of fuzzy transaction is computed with the membership for linguistic labels. The Fuzzy Support count across multiple transactions is computed for finding the fitness of the optimized function. The weight parameter is required to keep away from local minima that occur as a result of the magnitude of condense being higher than the magnitude of temporal fuzzy support.

In 2013 Binu Thomas et al. [6], proposed an algorithm whose concept is based on the web page weight, number of co-occurrence of web category and web access class, web category in sequence with maximum weight. The web pages are converted to fuzzy weights in a database. These web pages are sorted in descending order for selecting a web page with the maximum weight. The pages are considered, in classification rule as long as the product of the weights is greater that the support threshold. Those rules that have a confidence value greater than threshold confidence value are selected.

In 2013 Victoria Newport [7], the proposed method is able to identify and construct transactions according to the mining pattern. The set of transactions obtained are processed by a traditional pattern mining algorithm, which finds association rules of the form specified in the mining pattern along support and confidence greater than user's specified ones.

In 2012 S.Veeramalai et al. [8], an algorithm is proposed based on the Hash tree algorithm for frequent itemset and rule generation phase. The frequent itemsets are generated by combining all possible combinations of items known as candidate set. For the combination of the items the database is scanned multiple times. The support value of each itemset is counted and compared with the user-defined minimum support value, the item values that are greater are frequent itemsets. The Hash trees are used for storing all k-item sets and their counts. The frequent k-1 item sets, which have common parents, are considered for joining that avoids the

repeat scanning of Lk-1. The minimum support value is computed from the Hash trees that is 50% of the minimum support values. The proposed algorithm helps in faster access to the relevant information in the web mining search.

In 2011 Liu, et al. [9], proposed a method based on the basic Apriori rule mining algorithm. The paper aims to enrich the interpretation of the drafts of rule mining. Initially the database is pre-processed to obtain specific documents free from redundancy and local noisy information such as local caches and proxy servers. The algorithm discovers the rules based on the first support, second support and Rel-confidence parameters. The first support and second support parameters are entered by the user and the Rel-confidence parameter is computes by finding the maximum value among the meanest of the support of each item in the transaction.

In 2010 Tarek F. Gharib et al. [10], proposed an algorithm to maintain temporal frequent itemsets from the temporal transaction database. The incremental procedure is done by Sliding Window Filtering Technique. The algorithm updates the candidate itemsets with their support counts and thereby reduces scanning to find new itemsets with one database scan. It is designed to handle the problem of extending a partition done several times. The incremental procedure performs the pre-processing of the incremental mining. The algorithm checks for timestamp of the partition of the original database with the timestamp of the incremental database. If the values are equal updates and merges the transaction of the last partition of the original database with the incremental database. Otherwise the direct incremental database is considered that filters the candidates for scan reduction. The obtained candidates are transformed into temporal itemsets and sub-itemsets are generated from the scanned database.

In 2016 Ujwala Manoj Patil et al. [11], proposed Web usage mining is the method of extracting interesting patterns from Web usage log file. Web usage mining is a sub field of data mining uses various data mining techniques to generate association rules. Data mining techniques are used to generate association rules from transaction data. Most of the time transactions are boolean transactions, whereas Web usage data having quantitative values. To handle these real world quantitative data, we used fuzzy data mining algorithm for extraction of association rules from the quantitative web log file. To generate fuzzy association rules first we plan membership function. This member function is used to transform quantitative values in fuzzy terms. Experiments are performed on different support and confidence. The experimental output shows the performance of the algorithm with varied supports and confidence.

In 2013 R GEETHARAMANI et al. [12], proposed Web Usage mining is a technique used to recognize the user needs from the web log. Discovering hidden patterns from the logs is an approaching research area. Association rules play a key role in many web mining applications to find interesting patterns. However, it generates enormous rules that cause researchers give ample time and expertise to find the really interesting ones. This paper works on the server logs from the MSNBC dataset for the month of September 1999. This research aims at predicting the probable subsequent page in the usage of web pages listed in this data based on their navigating behavior by using Apriori prefix tree (PT) algorithm. The created rules were ranked based on the support, confidence and lift evaluation measures. The final predictions revealed that the interestingness of pages mainly depended on the support and the lift measure, whereas confidence assumed a uniform value among all the pages. It proved that the system guaranteed 100% confidence with the support of 1.3E−05. It revealed that the pages such as Front page, On-air, News, Sports and BBS attracted more interested subsequent users compared to Travel, MSN-News and MSN-Sports which were of less interest.

In 2014 Richa Soni et al. [13], proposed complete framework and findings in mining Web usage patterns from Web log files of a real Web site that has all the challenging things of real-life Web usage mining, containing evolving user profiles and external data showing an ontology of the Web data. We present an approach for finding and tracking evolving user profiles. We also define how the searched user profiles can be intensify with direct information need that is inferred from search queries take out from Web log data. Paper presents a knowledge discovery framework for the construction of Community Web Directories, a concept that we introduced in our recent work, applying customization to Web directories. In this context, the Web directory is seen as a thematic hierarchy and customization is realized by constructing user community models on the basis of usage data. We improve the clustering and probabilistic approaches presented in previous work and also shows a new algorithm that combine these two approaches. The resulting community models gain the form of Community Web Directories. The intended customization methodology is fixed on a distinct artificial and a general-purpose Web directory, indicating its potential value to the Web user. Web mining techniques look to take out knowledge from Web data. This paper give an overview of last and current work in the three main areas of Web mining research content, structure, and usage as well as emerging work in Semantic Web mining. Statistical testing and reliability analysis can be used effectively to conform quality for Web applications. To support this strategy, we consider Web usage and failure information from existing Web logs. The usage information is consider to make models for statistical Web testing. Optimizing parts before developing the system as a whole can help large company deploy efficient, geographically redundant Web infrastructures.

In 2012 Chhavi Rana et al. [14], proposed Web usage mining emphasis on techniques that could predict user behavior while the user interacts with the Web. It attempt to make sense of

the data discovered by the Web surfer's sessions or behaviors. There is way to give an overview of the state of the art in the research of web usage mining, while discussing the most related tools available in the sphere as well as the niche requirements that the current variety of tools lack. It will give an outlook on the existing tools, their specialized focus with respect to a useful objective and the need for a more comprehensive new entrant in this sphere in the light of the current scenario. In the end, the paper will be concluded by listing some challenges and future trends in this research area. The Overall the focus of the paper will be presenting a survey of the current developments in this area which is getting too much recognition from the web development arena.

In 2012 Stephen G. Matthews et al. [15], proposed that contributes the issues of finding fuzzy association rules that exhibit a temporal pattern. The other use of the 2- tuple linguistic presentation find fuzzy association rules in a temporal context, whilst maintaining the interpretability of linguistic terms. Iterative Rule Learning (IRL) with a Genetic Algorithm (GA) simultaneously make the rules and tunes the membership functions. The discovered rules are match with those from a old method of finding fuzzy association rules and output shows how the old method can loose information because rules fined at the intersection of membership function boundaries. New information can be extracted from the proposed approach by improving upon rules discovered with the traditional method and by discovering new rules.

III. SUMMARY OF LITERATURE REVIEW

Author	Method	Remark
Rajni Pamnani [2]	WUM, SUGGEST, incremental procedure	Dynamically create links to pages that have not yet been visited by a user. SUGGEST does not make use of any off-line component.
Kavita Sharma [3]	Web Content Mining, Web Structure Mining, Web Usage Mining	Show the superficial knowledge and comparison about data mining. The prefetching scheme increases the network traffic as well as the Web server's load.
Jiawei Han [4]	Frequent pattern tree, Mining Frequent pattern	FP-growth method, calculated its performance in comparison with many influential frequent pattern mining algorithms in large databases. SQL-based, highly scalable FP-tree structure, constraint-based mining of frequent patterns using FP-trees.
S.G. Matthews [5]	Fuzzy Apriori, CHC with 2-tuple linguistic representation	Show the problem of losing temporal fuzzy association rules on real-world Web log data. Lowering minimum support /confidence would conquered the problem of losing rules with traditional approaches, however, the number of rules increases, which is undesirable in association rule mining.
Binu Thomas [6]	Boolean Apriori algorithm, The Fuzzy Web Classification Algorithm (FWCA)	FWCA algorithm can find the longest possible frequent patterns in a single step by using the concepts of fuzzy weighted association. While the Apriori algorithm need many passes over the data to generate the rules.
S. Veeramalai [8]	Hash tree Algorithm, Apriori Hash Tree with Fuzzy, Modified Apriori Hash Tree	Crisp boundary problem in the joined algorithm and it is overcome by association Apriori hash tree fuzzy algorithm and the efficiency is increased in it. To develop the optimizing search application system.
Ujwala Manoj Patil [11]	FUZZY SET CONCEPTS, Fuzzy association rules	The given algorithm uses static membership functions to fuzzify the quantitative Web usage data along with predefined membership function. Possibility to loose some association rules, but in future it is possible to discover some interesting temporal association rules.
R. Geetharamani [12]	Apriori algorithm, Association rules	An Apriori - PT algorithm was utilized to mine the frequent pattern rules from which patterns were analyzed based on the rule consequents and Antecedents.
Richa Soni [13]	Semantic Web Mining, Web Usage Mining	The given methodology provides a promising research direction, where many new issues arise. An analysis Regarding the parameters of the community models, such As PLSA, is required.

IV. CONCLUSIONS

This paper presents a recent survey on the Web usage mining. With the rapid growth of the Web-based applications, there

has been a knee interest among the developers and users about the usage of the Web data and knowledge about different log parameters and analysis. The designing and improvement with the customer relation becomes better, thereby increasing the system performance. The mining of the Web data has been surveyed for the applications of various techniques such as association rule mining, clustering and fuzzy logic based applications.

REFERENCES

[1] Ujwala Patil and Sachin Pardeshi, "*A Survey on User Future Request Prediction: Web Usage Mining",* International Journal of Emerging Technology and Advanced Engineering, Vol. 2, Issue 3, pp. 121-124, 2012.

[2] Rajni Pamnani and Pramila Chawan, "*Web Usage Mining: A Research Area in Web Mining"*, pp. 2-3, 2012.

[3] Kavita Sharma, Gulshan Shrivastava and Vikas Kumar, "*Web Mining: Today and Tomorrow",* 3rd International Conference on Electronics Computer Technology, Vol. 1, pp. 399-403, 2011.

[4] Jiawei Han, Jian Pei, Yiwen Yin, *"Mining Frequent Patterns without Candidate Generation",* In Proceedings of the ACM-SIGMOD International Conference on Management of Data, pp. 1–12, 2000.

[5] Stephen G. Matthews, Mario A. Gongora, Adrian A. Hopgood and Samad Ahmadi, *"Web usage mining with evolutionary extraction of temporal fuzzy association rules",* Knowledge-Based Systems, Vol.54, pp. 66-72, 2013.

[6] Binu Thomas and G. Raju, *"A Novel Web Classification Algorithm Using Fuzzy Weighted Association Rules",* Hindawi Publishing Corporation, Vol.2013, pp. 1-10, 2013.

[7] Victoria Nebot and Rafael Berlanga, *"Finding association rules in semantic web data",* Knowledge-Based Systems, Vol.25, pp. 51-62, 2012.

[8] S. Veeramalai, N. Jaisankar and A. Kannan, *"Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy",* International Journal of Computer Science & Information Technology, Vol.2, No. 4, pp. 60-74, 2010.

[9] Liu Jian and Wang Yan-Qing, *"Web Log Data Mining Based on Association Rule",* Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Vol.11, pp. 1855-1859, 2011.

[10] Tarek F. Gharib, Hamed Nassar, Mohamed Taha and Ajith Abraham, *"An efficient algorithm for incremental mining of temporal association rules",* Data and Knowledge Engineering, Vol.69, pp. 800-815, 2010.

[11] Ujwala Manoj Patil and Prof. Dr. J. B. Patil, "MINING FUZZY ASSOCIATION RULES FROM WEB USAGE QUANTITATIVE DATA" CSITY, SIGPRO, AIFZ, NWCOM, DTMN, GRAPHHOC, pp. 89–98, 2016.

[12] R GEETHARAMANI, P REVATHY and SHOMONA G JACOB, *"Prediction of users webpage access behaviour using association rule mining"* Sadhana, Vol. 40, part 8, pp. 2353–2365, 2015.

[13] Richa Soni and Gurpreet Kaur, *" Web Usage Mining: Personalization of Web Usage Data",* International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 2, 2014

[14] Chhavi Rana, *"A Study of Web Usage Mining Research Tools",* Int. J. Advanced Networking and Applications, Vol. 03 Issue:06, pp. 1422-1429, 2012.

[15] Stephen G. Matthews, GONGORA, Mario A., HOPGOOD, Adrian A. and AHMADI, Samad, *"Temporal Fuzzy Association Rule Mining with 2-tuple Linguistic Representation",* IEEE International Conference on Fuzzy Systems, pp 1-8, 2012.

Authors Profile

Ujwala M. Patil has completed her master of technology in Computer Engineering, Dr. Babasaheb Ambedkar Technological University, Lonere, Raigad, Maharashtra, India in 2007 and pursuing her Ph.D. in Computer Engineering from North Maharashtra University, Jalgaon, Maharashtra, India. She is working as an associate professor in the Computer Engineering Department at R.C. Patel Institute of Technology, Shirpur (Maharashtra), India. She has 13 years of teaching experience. Her research interests lie in machine learning, web usage mining, data mining, and their applications.

Pramod Pardeshi has completed his bachelor of engineering in Computer Engineering from North Maharashtra University, Jalgaon, Maharashtra, India in 2010 and pursuing his master of engineering in Computer Engineering from North Maharashtra University, Jalgaon, Maharashtra, India.

