MINING FUZZY ASSOCIATION RULES FROM WEB USAGE QUANTITATIVE DATA

Ujwala Manoj Patil and Prof. Dr. J. B. Patil

Department of Computer Engineering, R.C.P.I.T., Shirpur, Maharashtra, India.
patilujwala2003@gmail.com
jbpatil@hotmail.com

ABSTRACT

Web usage mining is the method of extracting interesting patterns from Web usage log file. Web usage mining is subfield of data mining uses various data mining techniques to produce 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 consists of quantitative values. To handle these real world quantitative data we used fuzzy data mining algorithm for extraction of association rules from quantitative Web log file. To generate fuzzy association rules first we designed membership function. This membership function is used to transform quantitative values into fuzzy terms. Experiments are carried out on different support and confidence. Experimental results show the performance of the algorithm with varied supports and confidence.

KEYWORDS

Web Usage mining, Data mining, Fuzzy association rules, Web log file, Fuzzy term.

1. INTRODUCTION

With the continued increase in the usage of the World Wide Web (WWW), Web mining has been established as an important area of research. The WWW is a vast repository of unstructured information, in the form of interrelated files; those are distributed on several Web servers over wide geographical regions. Web mining deals with the discovering and analyzing the valuable information from the WWW. Web usage mining focuses on discovery of the potential knowledge from the browsing patterns of users to find the correlation between the pages on analysis.

Mining is of three types: Data Mining, Text Mining and Web Mining. There are many challenging problems in Data, Text, and Web Mining Research. The mining data may be either structured or unstructured. Data Mining deals with structured data organized in a database whereas text mining deals with unstructured data. Web mining data handles the combination of structured and unstructured data. Web Mining uses data mining as well as text mining techniques and its distinctive approaches. Web data mining is the application of data mining techniques to discover interesting and potentially useful knowledge from Web data. Web hyperlink structure or Web log data or both are used by Web data mining process.

David C. Wyld et al. (Eds) : CSITY, SIGPRO, AIFZ, NWCOM, DTMN, GRAPHHOC - 2016
pp. 89–98, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60408

There are many types of data that can be used in Web Mining [1, 2].

1.1 Web Content

The data actually present in the Web pages which conveys information to the users. The contents of a Web page may be varied e.g. text, HTML, audio, video, images, etc.

1.2 Web Structure

The organization of the Web pages connected through hyperlinks i.e. various HTML tags used to link one page to another and one Web site to another Web site.

1.3 Web Usage

The data that reflect the usage of Web collected on Web servers, proxy server, and client browser with IP address, date, time etc.

1.4 Web User Profile

The data that provides demographic information about users of the Web sites, i.e. user registration data and customers profile information.

World-wide-Web applications have grown very fast and have made a significant impact on computer systems. Among them, Web browsing for useful information may be most commonly seen. Due to its incredible amounts of use, efficient and effective Web retrieval has thus become a very important research topic in this field.

Figure 1 shows the step wise procedure for Web usage mining process.

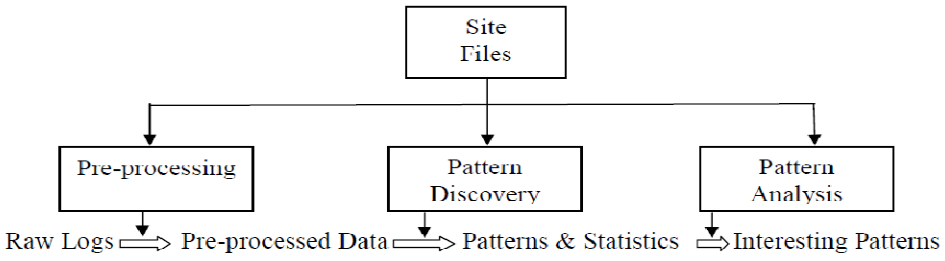

Figure 1. Web Usage Mining Process

The general process of Web usage mining includes [3]

1. Resource collection: Process of extracting the task relevant data (e.g. access logs of HTTP servers),
2. Information pre-processing: Process of Cleaning, Integrating and Transforming of the result of resource collection,
3. Pattern discovery: Process of uncovered general patterns in the pre-process data and
4. Pattern analysis: Process of validating the discovered patterns.

There are different Web mining techniques [1], used for efficient and effective Web retrieval. Web usage mining is one of the ways for the same. Web-usage mining emphasizes on the automatic discovery of user access patterns from Web servers [1, 2]. In the past, several Web-mining approaches for finding sequential patterns and user interesting information from the World Wide Web were proposed [1, 2, 4].

Real world transactions are commonly seen with quantitative values known as boolean transactions [5]. A boolean association involves binary attributes; a *generalized* association involves attributes that are hierarchically related and a *quantitative* association involves attributes that can take on quantitative or categorical values. For example, assume whenever customers in a supermarket buy bread and butter, they will also buy milk. From the transaction of the supermarkets, an association rule can be mined out as "Bread and Butter →Milk". Most of the previous study focused on such type of boolean transaction data. Transaction data in real-world applications usually consist of quantitative values. Designing a sophisticated data-mining algorithm which will handle real-world applications data presents a challenge to data mining researchers.

Exhaustive research has been done in Web mining. There are many more techniques used to find association between Web pages. But instead of only page sequence if we consider page view time while accessing the Web page then the Web log sequence can be seen as quantitative data.

Fuzzy logic, which may be viewed as an extension of traditional logical systems, provides an effective conceptual framework for dealing with the problem of knowledge representation in an environment of uncertainty and imprecision [6, 7, 8, 9].

Fuzzy set theory was first introduced by Zadeh in 1965 [6].

Fuzzy set theory is being used more and more frequently in intelligent systems because of its simplicity and similarity to human reasoning [5, 6, 7, 8, 9]. The theory has been applied in fields such as manufacturing, engineering, control, diagnosis, and economics, among others [6, 8].

Here first, we applied fuzzy concept to Web usage log data to find fuzzy labels and then applied Apriori algorithm to find interesting association rules.

The remaining parts of this paper are organized as follows: Basic Apriori algorithm is reviewed in Section 2. Fuzzy set concepts are reviewed in Section 3. Proposed algorithm is described in Section 4. Experimental results are shown in Section 5. Conclusions and future work is described in the last Section 6.

2. REVIEW OF BASIC APRIORI ALGORITHM

Agrawal et.al developed several mining algorithms based on the concept of large itemsets to find association rules between transaction data [4]. Sequential mining process is based on two phases a, phase one is used to generate number of candidate itemsets and then its frequency is counted by scanning the transaction data. The qualifying itemsets i.e. which are equal to or above predefined threshold value called as min support are called as large itemsets. Initially large-1 itemsets are generated; these large-1 itemsets are combined to form large-2 itemsets and so on. This process is repeated until all large itemsets are had been found. In second phase, all association rules are

formed against large itemsets. Then each association rule is checked with min confidence. Qualifying association rules were output as set of association rules.

Input- L_1= {large-1 sequences}
Output- maximal sequences in $\cup_k L_k$
For (k=2;L_{k-1};k++) do
begin
C_k=New candidates generated from L_{k-1}
For each user-sequence c in the database do
Increment the count of all candidates in C_k that are contained in c.
L_k = Candidates in C_k with minimum support.
end

3. REVIEW OF FUZZY SET CONCEPTS

Formally, the process by which individuals from a universal set X are determined to be either members or non-members of crisp set can be defined by a characteristic or discrimination function [5, 6]. For a given crisp set A, this function assigns a value $\mu_A(x)$ to every x ∈ X such that

$$\mu_A(x)=\begin{cases}1 \; if \; and \; only \; if \; x \in A \\ 0 \; if \; and \; only \; if \; x \notin A\end{cases}$$

Thus, this function maps elements of the universal set to the set containing 0 and 1. This kind of function can be generalized such that the values assigned to the elements of the universal set fall within specified ranges, referred to as the membership grades of these elements in the set. Higher the value denotes better degrees of the set membership. Such a function is called membership function $\mu_A(x)$, by which fuzzy set A is usually defined. This function is represented by

$$\mu_A: X \rightarrow [0, 1],$$

Where [0, 1] denotes the interval of real numbers from 0 to 1, inclusive.

A special notation is often used in the literature to represent fuzzy sets, Assume that x_1 to x_n are the elements in fuzzy set A, and μ_1 to μ_n are their grades of membership in A, A is usually represented as follows:

$$A= \mu_1/x_1 + \mu_2/x_2 + \ldots.. + \mu_n/x_n$$

3.1 Operations on Fuzzy Sets

Following are the basic and commonly used operations on fuzzy sets as proposed by Zadeh [5].

3.1.1 Complementation

The complementation of a fuzzy set A is denoted by ¬A, and the membership function of ¬A is given by:

$$\mu_{\neg A}(x)= 1 - \mu_A(x), \; \forall x \in X$$

3.1.2 Union

The union of fuzzy sets A and B is denoted by A∪B, and the membership function of A∪B is given by:

$$\mu_{A\cup B}(x) = \max\{\mu_A(x), \mu_B(x)\} \quad \forall x \in X$$

3.1.3 Intersection

The union of fuzzy sets A and B is denoted by A∩B, and the membership function of A∩B is given by:

$$\mu_{A\cap B}(x) = \min\{\mu_A(x), \mu_B(x)\} \quad \forall x \in X$$

4. PROPOSED ALGORITHM

```
Input: Pre-processed dataset
Output: Set of fuzzy association rules
Get initial membership functions, support value and confidence value
Divide dataset into partitions
Set n to number of partitions to be processed
repeat
        Transfer quantitative values into fuzzy terms with fuzzy values
        Calculate the counts of fuzzy terms
        repeat
                for each fuzzy term
                        Generate the candidate set by counting of each fuzzy term
                end for
                for each fuzzy term
                        if count ≥ min support
                                generate large itemsets
                        end if
                end for
                join the large itemsets
        until large itemsets = = NULL
until n=0
repeat
for each large itemsets
        if confidence ≥ min confidence
                Construct association rule
        end if
end for
merge all association rules
until n=0
```

5. EXPERIMENTAL RESULTS

The experimental results are derived from United States Environmental Protection Agency (EPA). The EPA dataset contains a 24-hour period of Hypertext Transfer Protocol (HTTP) requests to a Web server [11]. The EPA dataset has HTTP request from 23:53:25 EDT 29th August 1995 to 23:53:07 30th August 1995. The EPA dataset has total 47748 requests: 46014 GET requests, 1622 POST requests, 107 HEAD requests and 6 invalid requests. Table 1 shows a sample of records from the EPA dataset after cleaning and preprocessing.

Table 1. Input Transactions

TID	Pages with view time in seconds
0	(1,26); (2,260); (3,120); (4,430)
1	(6,220); (7,86); (8,9); (9,101); (10,320)
2	(6,22);(7,520);(8,17)
3	(13,190);(14,74)
4	(15,6);(16,140);(17,133);(18,261)
5	(6,136);(20,880)

The quantitative value of each transaction is transformed into fuzzy set according to the membership functions defined in figure 2.

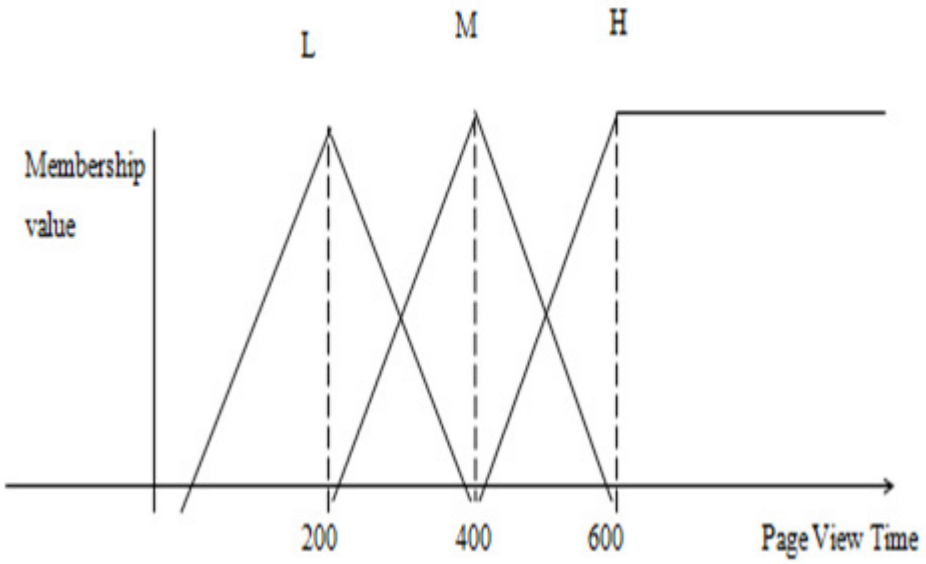

Figure 2. A Triangular Membership Functions for Page View Time

All the transactions from table 1 are converted in fuzzy terms (shown in table 2) using the membership functions defined in figure 2.

Table 2. Input Transactions transformed into the fuzzy sets

TID	Fuzzy set
0	$(\frac{1.0}{1.low})$; $(\frac{0.7}{2.low}+\frac{0.3}{2.medium})$; $(\frac{1.0}{3.low})$; $(\frac{0.85}{4.medium}+\frac{0.15}{4.high})$;
1	$(\frac{0.9}{6.low}+\frac{0.1}{6.medium})$; $(\frac{1.0}{7.low})$; $(\frac{1.0}{8.low})$; $(\frac{1.0}{9.low})$; $(\frac{0.4}{10.low}+\frac{0.6}{10.medium})$
2	$(\frac{1.0}{6.low})$;$(\frac{0.4}{7.medium}+\frac{0.6}{7.high})$;$(\frac{1.0}{8.low})$
3	$(\frac{1.0}{13.low})$;$(\frac{1.0}{14.low})$
4	$(\frac{1.0}{15.low})$;$(\frac{1.0}{16.low})$;$(\frac{1.0}{17.low})$;$(\frac{0.7}{18.low}+\frac{0.3}{18.medium})$
5	$(\frac{1.0}{6.low})$;$(\frac{1.0}{20.high})$

Calculate the scalar cardinality of each fuzzy term in given transactions as the count value. The counts for all fuzzy terms are shown in table 3. This fuzzy terms set is called candidate -1 fuzzy term set.

Table 3. Counts for all fuzzy terms

Fuzzy term	Count	Fuzzy term	Count
1.low	1.0	10.medium	0.6
2.low	0.7	7.medium	0.4
2.medium	0.3	7.high	0.6
3.low	1.0	13.low	1.0
4.medium	0.85	14.low	1.0
4.high	0.15	15.low	1.0
6.low	2.9	16.low	1.0
6.medium	0.1	17.low	1.0
7.low	1.0	18.low	0.7
8.low	2.0	18.medium	0.3
9.low	1.0	20.high	1.0
10.low	0.4		

The candidate-1 fuzzy term set is checked against predefined minimum support value assume that the support value is 1.0. The qualified fuzzy terms are called as large-1 fuzzy terms shown in table 4.

Table 4. The set of large-1 fuzzy terms

Fuzzy term	count	Fuzzy term	count
1.low	1.0	13.low	1.0
3.low	1.0	14.low	1.0
6.low	2.9	15.low	1.0
7.low	1.0	16.low	1.0
8.low	2.0	17.low	1.0
9.low	1.0	20.high	1.0

As large-1 fuzzy term set is not null, join large-1 fuzzy terms to generate candidate-2 fuzzy terms. Then we check candidate-2 fuzzy terms with minimum support to find large-2 fuzzy term set. Hence we repeat the process of join and generate till we get the largest fuzzy term set null. Finally we combined all partitions largest fuzzy term sets to form association rules. Then each association rule is checked with min confidence. Qualifying association rules were output as set of association rules. For example visitors who viewed page 7 for low amount time also viewed page 8 for low amount of time.

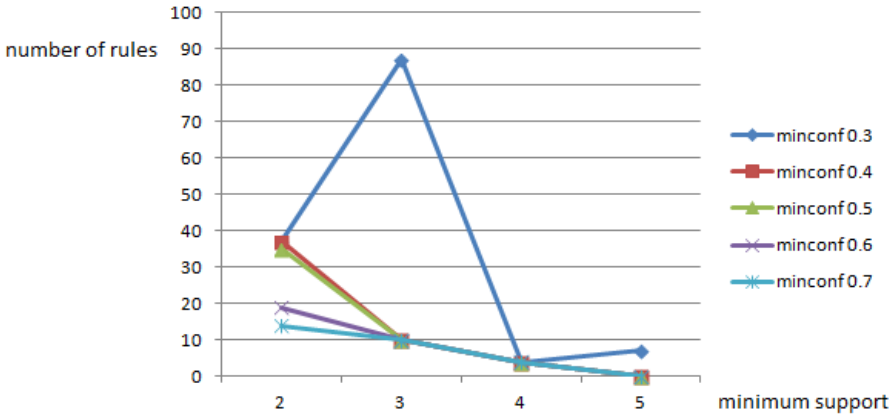

Figure 3. The relationship between numbers of association rules and minimum support values.

Experiments were conducted with varied support and confidence. From figure 3, it is clear that the number of association rules decreased along with the increase in minimum support values.

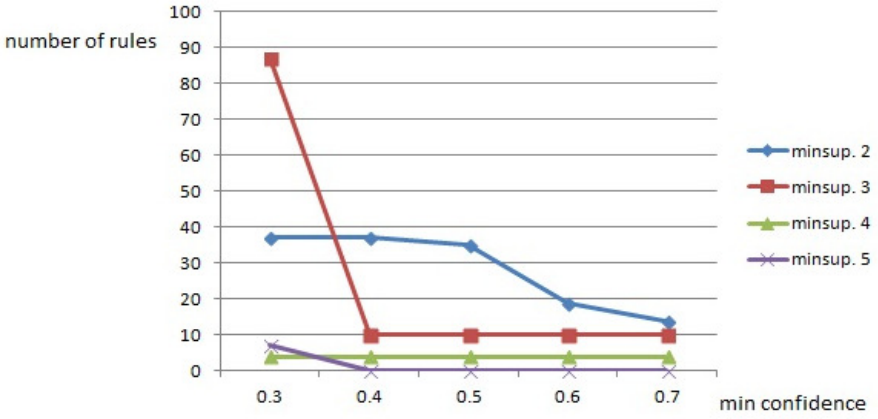

Figure 4. The relationship between numbers of association rules and minimum confidence values.

From figure 4, it is easily seen that the number of association rules decreased as the increase in minimum confidence values. When we observe the curves of figure 3 and figure 4, it is clear that the curves for larger minimum support values were smoother than smaller minimum support values; it means that minimum confidence value had a larger effect on number of association rules when smaller minimum support values were used.

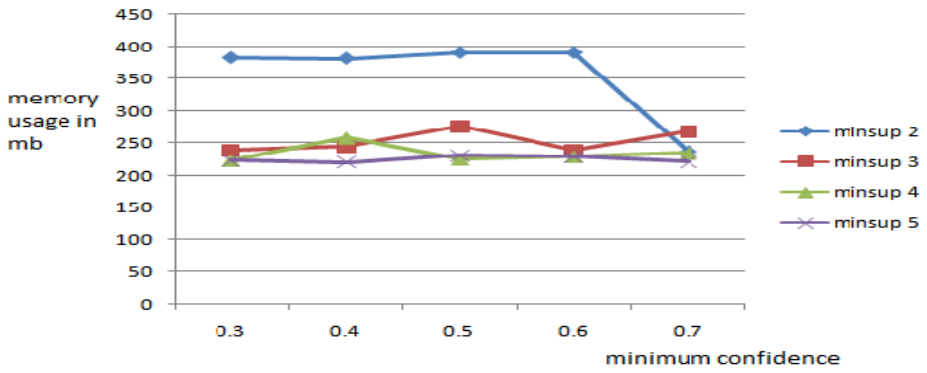

Figure 5. The Memory Usage Statistics with varied minimum support and confidence.

From figure 5, it is observed that the memory utilization is high when the minimum support is low and memory utilization gets decreased as minimum support values get increased.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a generalized fuzzy data mining algorithm to extract interesting patterns. The proposed algorithm uses static membership functions to fuzzify the quantitative Web usage data along with predefined membership function. We also use predefined support and confidence. In this paper we divided whole database into different partitions based on hours. Each hour partition, we apply separately fuzzy mining algorithm to extract association rules. Finally all hours association rules combined to declare total number of rules for given database. There is possibility to lose some association rules, but in future, we will try to attempt to discover some interesting temporal association rules based on this partition.

REFERENCES

[1] J. Srivastava, R. Cooley, M. Deshpande, and P. -N. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," ACM SIGKDD Explorations, Vol. 1, no. 2, pp. 1-12, Jan. 2000.

[2] R. Cooley, B. Mobasher and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web," Ninth IEEE International Conference on Tools with Artificial Intelligence, pp. 558–567, Nov. 1997.

[3] R. Agrawal and R. Srikant, "Mining Sequential Patterns," Proceedings of the 11th Conference on Data Engineering, Taipei Taiwan, IEEE Computer Society Press, pp. 3-14, 1995.

[4] R. Cooley, B. Mobasher, and J. Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns," Knowledge and Information Systems, Vol. 1, No. 1, pp. 5-32, 1999.

[5] T.P. Hong, C.S. Kuo, and S.C. Chi, "Trade-off between computation time and number of rules for fuzzy mining from quantitative data," International Journal of Uncertainty, Fuzziness & Knowledge-Based Systems, vol. 9, no 5, 2001, pp. 587–604.

[6] L.A. Zadeh, "Fuzzy sets," Information Control, vol. 8, Issue 3, June 1965, pp. 338–353.

[7] T.P. Hong, and C.Y. Lee, "An overview of mining fuzzy association rules," In: H. Bustince, F. Herrera, J. Montero (eds.) Studies in Fuzziness and Soft Computing, vol. 220, pp. 397–410. Springer Berlin/Heidelberg (2008)

[8] L.A. Zadeh, "Knowledge representation in fuzzy logic," IEEE Transactions on Knowledge and Data Engineering, vol. 1, no. 1, March 1989, pp. 89-100.

[9] T.P. Hong, K.Y. Lin, and S.L. Wang, "Mining linguistic browsing patterns in the world wide Web," Soft Computing – A Fusion of Foundations, Methodologies and Applications, vol. 6, no 5, 2002, pp.329–336.

[10] T.P. Hong, and C.Y. Lee, "Induction of fuzzy rules and membership functions from training examples," Fuzzy Sets and Systems, vol. 84, 1996, pp. 33-47.

[11] Stephen G. Matthews, M.A. Gongora, A.A. Hopgood, and S. Ahmadi, "Web usage mining with evolutionary extraction of temporal fuzzy association rules," Knowledge based Systems, vol. 54, 2013, pp. 66-72.

AUTHORS

Jayantrao B. Patil has completed master of technology in computer science & data processing from IIT, Kharagpur and Ph.D. in computer engineering from North Maharashtra University, Jalgaon, Maharashtra, India. He is working as a principal at R. C. Patel Institute of Technology, Shirpur (Maharashtra), India. His area of research is web catching and web prefetching, web data mining, text watermarking, web usage mining, web personalization, semantic web mining and web security. He is a life member of Indian Society for Technical Education (ISTE), Computer Society of India (CSI), the member of Institute of Engineers (IE), India and the senior member of International Association of Computer Science and Information Technology (IACSIT), Singapore**.**

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.