International Journal of Fuzzy Logic and Intelligent Systems, vol. 2, no. 3, December 2002 pp. 204~209

Fuzzy Web Usage Mining for User Modeling

Jae-Sung Jang*, Sung-Hae Jun and Kyung-Whan Oh****

*** S/W Department, LG Electronics Inc.**
150-010 Seoul, Korea
**** Department of Computer Science, Sogang University**
121-742 Seoul, Korea

Abstract

The interest of data mining in artificial intelligence with fuzzy logic has been increased. Data mining is a process of extracting desirable knowledge and interesting pattern from large data set. Because of expansion of WWW, web data is more and more huge. Besides mining web contents and web structures, another important task for web mining is web usage mining which mines web log data to discover user access pattern. The goal of web usage mining in this paper is to find interesting user pattern in the web with user feedback. It is very important to find user's characteristic for e-business environment. In Customer Relationship Management, recommending product and sending e-mail to user by extracted users characteristics are needed. Using our method, we extract user profile from the result of web usage mining. In this research, we concentrate on finding association rules and verify validity of them. The proposed procedure can integrate fuzzy set concept and association rule. Fuzzy association rule uses given server log file and performs several preprocessing tasks. Extracted transaction files are used to find rules by fuzzy web usage mining. To verify the validity of user's feedback, the web log data from our laboratory web server.

Key Words : User Modeling, Fuzzy Process, Fuzzy Web Usage Mining, MinSup and MinConf

1. Introduction

Recently, interest of data mining in artificial intelligence is increasing. Data mining is the process of extracting desirable knowledge and interesting pattern. Because of development of WWW, web data is more and more huge. Goal of web usage mining is to find interesting user pattern in the web. In the environment of EC(Electronic Commerce), finding users characteristics is very important. It is possible to recommend product and send e-mail to user by extracted users characteristics. Users profile can be extracted by web usage mining. In our research, we concentrate on finding association rules and verify validity of association rules. The proposed algorithm integrates fuzzy set concepts and association rule. Fuzzy association rules use given server log file. There are several preprocessing tasks are performed. Extracted transaction files are used to find rules by fuzzy association rule. To verify validity, users feedback is demanded. Goal of users feedback is to reach high level of users satisfaction. This paper is organized as follows. In the following section, we will describe fuzzy association rule in web mining. We will account for fuzzy association rule considering user feedback in section 3. In section 4, the experimental results will be given. In this experiments, the number of extracted association rules are analyzed. Using users feedback, degree of satisfaction is extracted. We improved performance of association rule by learning of users feedback. We will give a brief conclusion in section 5.

This paper is supported by Brain Science project sponsored by Korea Ministry of Science and Technology and partially supported by Industrial Technology Institute at Sogang University.

2. Fuzzy Web Usage Mining

2.1 Fuzzy Process in User Transaction

The relation between the user's successive access path is always regarded as a sequential association relation. Discovering the association rule by using the concepts of minimum support and confidence is used. But the accuracy of prediction using these approaches is not much better than other algorithms. One of the possible reasons for this is the lack of a fuzzy concept in the rule. For example, there are two rules: "If A and B then C" with support equal to 80% and confidence equal to 50%, and "If A and B then D" with support equal to 79% and confidence equal to 49%. If there is an active user session with the first two accesses equal to A and B, then the first rule is chosen. However, we can see that the second rule is as strong as the first rule. Therefore, there should be an equal choice for following either the first and second rule for finding a correct match. In order to make rules more sensible and reasonable, fuzzy process is introduced in our paper.

2.2 Web Usage Mining

Web data has been enormous according as internet increased. This brings to difficulty of information searching, knowledge creating, and user modeling. Therefore, we need an automatic tool to take information from web for solving these problems. The technique of user's pattern classification is

needed for the automatic tool. This is called a web mining. The web mining is a challenging task that searches for web access patterns, web structures, and the regularity and dynamics of web contents. This has three types[1]. There are content mining, structure mining and usage mining in web data mining. First, web content mining is to analyze contents in web pages. Second, web structural mining is to analyze summarized structure about web site and web page. The last, web usage mining is to find access pattern of connected user from web log files. In this section, we discuss web usage mining. User's transaction and cookies of server are used for web usage mining. Our web usage mining process has three steps which are preprocessing, pattern discovery and pattern analysis. In preprocessing step, we clean raw web log data for next analysis. The data size after this step is smaller than origin data. Next, in pattern discovery step, Some methods as statistical analysis, association analysis, clustering, classification and sequential analysis are used. The last, in pattern analysis step, statistics, visualization, usability analysis and database querying are used for understanding each user. In this paper, the application of web usage mining can be used for finding user's behavior.

2.3 Fuzzy Association Rule Mining

The association rules are found from items and transactions[8]. Let $I=\{i_1, i_2, \ldots, i_m\}$ be a item set and D be a database transaction set where each transaction T is a set of items such that $T \subseteq I$. Also, let A be a item set. A T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset I$, $B \subset I$, and $A \cap B = \emptyset$. The rule $A \Rightarrow B$ holds in the transaction set D with support s, where s is the percentage of transactions in D that contain $(A \cap B)$. This is taken to be the probability, $P(A \cap B)$. The rule $A \Rightarrow B$ has confidence c in the transaction set D if c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability, $P(B|A)$. That is, $support(A \Rightarrow B) = P(A \cap B)$ and $confidence(A \Rightarrow B) = P(A \mid B)$. Rules that satisfy both a minimum support(MinSup) threshold and a minimum confidence(MinConf) threshold are called strong. By convention, we write support and confidence values so as to occur between 0% and 100%, rather than 0 to 1.0. In our research, the fuzzy process is used for optimal association rules. The fuzzy association rule algorithms proposed by Gyenesei is applied in this[3]. Given a database $D=\{d_1, d_2, \ldots, d_n\}$ with attribute I and the fuzzy sets associated with attributes in I, we want to find out some interesting. The following form for fuzzy association rule is used. If $X=\{x_1, x_2, \ldots, x_p\}$ is $A=\{f_1, f_2, \ldots, f_p\}$ then $Y=\{y_1, y_2, \ldots, y_p\}$ is $B=\{g_1, g_2, \ldots, g_p\}$, where f_i is the element of fuzzy sets related to attribute x_i , g_j is the element of fuzzy sets related to attribute y_j. And X and Y are itemsets. A and B contain the fuzzy sets associated with the corresponding attributes in X and Y. As in the binary association rule, 'X is A' is called the antecedent of the rule while 'Y is B' is called the consequent of the rule. The concept of fuzzy set is used for representation of quantitative and binary property in web mining through fuzzy association rule. The binary association rule is changed into the fuzzy association rule in this paper. First of all, the binary association rule is considered. In item set $I=\{i_1, i_2, \ldots, i_n\}$, each element of this set is a pair $\langle x, v \rangle$. The x is attribute and v is a value of x. $D=\{d_1, d_2, \ldots, d_n\}$ is a database. The d_i is ith tuple of D. This is a quantitative attribute. The following definition is needed to show attributes by fuzzy set[9]. The $F_{ik}=\{f^1_{ik}, f^2_{ik}, \ldots, f^j_{ik}\}$ is fuzzy set which is related with index ik, where f^j_{ik} is jth fuzzy set of F_{ik}. Preprocessing results of web log data are analyzed to transfer the useful information of web pages into fuzzy set using access frequency. The transaction file after data processing has quantitative association. We get a fuzzy value from them[5].

2.4 Criterion of Fuzzy Web Usage Mining

To extract the fuzzy association rule, first, the transaction support of all sets of items are computed. The item set with minimum support is frequent item set. A fuzzy support value has the number of records supporting the item set and their degree of support. The fuzzy confidence value is the measure of degree of support given by records. Therefore, confidence is used for us to estimate the interestingness of the generated fuzzy association rules. In this paper, the formula proposed by Gyenesei is adapted[3]. The Support and confidence may be useful for many applications. However, the support-confidence framework can be misleading in that it may identify a rule $A \Rightarrow B$ as interesting when the occurrence of A does not imply the occurrence of B. So we consider an alternative framework for finding interesting relationships between data item sets based on correlation. A correlation value is used for this framework as[4].

$$Corr_{A,B} = \frac{P(A \cap B)}{P(A)P(B)} \quad (1)$$

If the computing value of equation (1) is less than 1, then the occurrence of A is negatively correlated with the occurrence of B. If the computing value is greater than 1, then A and B are positively correlated, meaning the occurrence of one implies the occurrence of the other. If the computing value is equal to 1, then A and B are independent and there is no correlation between them. Equation (1) is equivalent to $P(B|A)/P(B)$, which is also called the lift of the association rule $A \Rightarrow B$.

3. Fuzzy Association Rule for User Modeling

3.1 Personalization and User Modeling

The personalization is to analyze each user using results of web usage mining and to support different service for each user[6]. It is easily to collect and systematize information of

user accessing web site. So, each user is serviced characteristically. Any user searches web site effectively and saves browsing time to get satisfied web site. User modeling is needed for constructing personalization. The user modeling is a learning process of user behaviors for effective user management. There are many fields in user modeling. Intelligent user interface is studied by many study group of user modeling. Conclusively, we extract interesting rule from user behaviors of the internet by user modeling method. The user modeling has a two types. These are implicit and explicit modeling. This is segmented by users conscious intervention. Usually all of two user modeling have been used. The implicit user modeling has not user's conscious intervention. And the explicit user modeling has user's conscious intervention in user modeling process. We change contents and structures of web page for showing to user using web browser of user modeling. And use intelligent interface to show recommended web page for user according to his interest. In this paper, explicit user modeling is used because of its convenience and construct user modeling by users interest. So, suitable user modeling can be possible.

3.2 Learning from user's satisfaction and feedback

User feedback learning process for user profile is represented as[7],

$$P = P + f \times U \ (\text{if } f > 0) \text{ and } P = 0 \ (\text{if } f \leq 0) \qquad (2)$$

where, P is learned profile of user, U is user interest, and f is a feedback. User profile is updated when the value of this expression is positive. When they have affirmative feedback of user, the weights of profile are increased. Elsewhere, they are changed to 0. User's implicit feedback is shown following table. To parallel, explicit feedback from user with extracting implicit feedback from experiment is gotten. Also we take feedback results from user and use results of from 0 to 1 for user input as feedback of profile learning. The results of fuzzy usage mining are applied to service a interesting rule.

Table 1. Example of implicit feedback

Partition	Feedback
buy recommended goods	2
click banner ad.	1.5
visit recommended site	1.2
delete recommended mail	0

3.3 Procedure Design

In our fuzzy web usage mining for user modeling, a preprocessing takes precedence over the others. Web log data are highly susceptible to noise, missing, and inconsistent data due to their typically huge size. There are a number of data preprocessing techniques. They are data cleaning, data integration, data transformations, and data reduction. For these, we perform data cleaning, user identification, session identification, path completion and transaction identification in our paper. Next, Fuzzy process is used for web usage mining. These two works are shown in Figure 2.

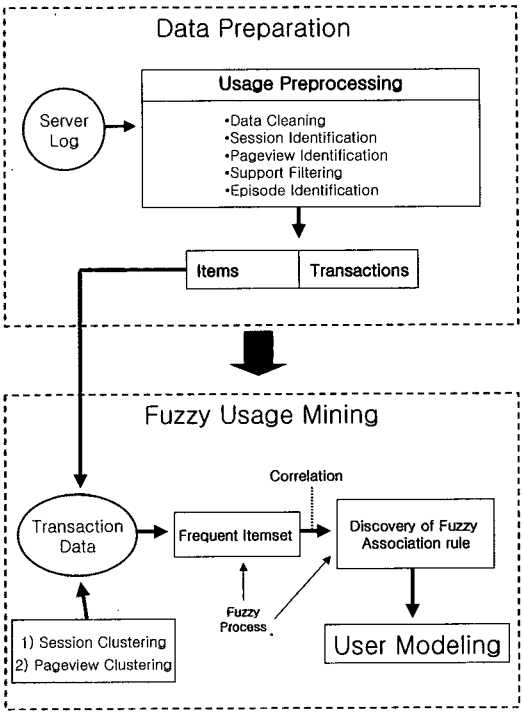


Fig. 1. Procedure Design

4. Experimental Results

To verify the advantages of association rule extracting with user feedback, fuzzy web mining technique is used. Real-world data form our web server are used for this experiment.

4.1 Experimental Design

In our experiment, user accessing log file in web server is analyzed for fuzzy usage mining by fuzzy association rule. The web log data from our laboratory web server system(ailab.sogang.ac.kr) is used for our experiment. The work is to analyze user connected behavior to our web server and request user feedback from serviced association rules. Figure 1 is showed the web site structure of our web server structure.

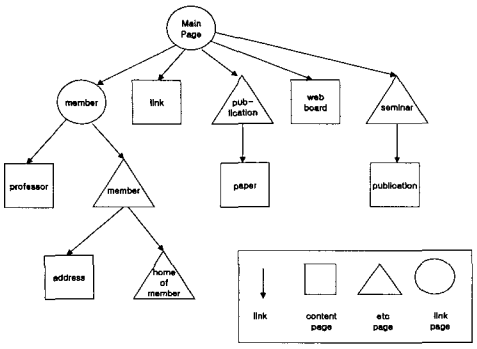


Fig. 2. Web site structure of our web server

4.2 Data Preparation

The log data of user connecting to our server during a month is used for experiments. This has been saved in our server(/usr/local/apache/logs). The useful data from this directory is extracted using preprocessing. This useful data is used for fuzzy web mining. In our paper, useful data has IP address, accessed time, accessed page, page size, and previous accessed information. The amount of user reference transactions are extremely large. The goal of data preparation is to extract the set of page view for analyzing user profile. In our experiment, we use the preparation technique by Cooley[2].

4.3 Extraction of association rules

We get some users who have useful transactions by preprocessing. They must have minimum transactions for experiment. The minimum transaction size is determined subjectively. In this paper, it is determined by 500. Table 2 shows the satisfied user IP by preprocessing.

Table 2. Extracted IP with meaningful transaction

211.44.66.97	163.239.235.131
163.239.135.27	164.124.106.101
165.132.122.31	203.230.220.91
202.30.128.20	163.239.45.37
163.239.35.249	163.239.5.202
163.239.130.20	150.150.55.22
143.248.119.254	192.35.17.26
128.134.29.9	210.98.149.73

The size of rule by fuzzy association rule from extracting result is changed using support and confidence. The result is shown in Fig. 3. In this, the adjusted relation from extracted association rules with user's required minimum confidence can be shown. The confidence is in inverse proportion to the number of association rules.

The relation of the number of association rules with minimum confidence is shown in Fig. 4. It shows that the minimum confidence is in inverse proportion to the number of association rules.

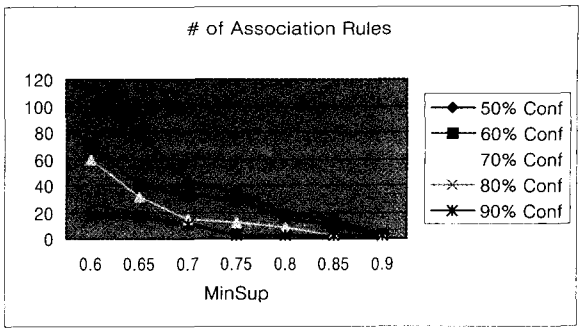


Fig. 3. Relation MinSup with the number of association rules

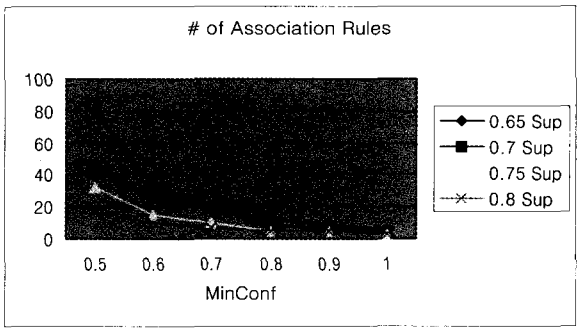


Fig. 4. Relation MinConf with the number of association rules

We know correlation of extracted association rule from Fig. 3 and 4. It is important that we can find most suitable association rule for user behavior understanding.

4.4 Pertinence evaluation of association rules by user feedback

A pertinence evaluation of extracted rules is needed for fuzzy web usage mining. Too many extracted rules may have insignificant rules. On the contrary, if so few association rule is extracted, the significant user intension are not included in them. Therefore, the suitable number of association rules must be decided by user feedback. They are stored in user profile by learning. User profile of learned association rules are used to determine necessary rules. According to preprocessed log data of our lab sever, 15 meaningful users are used for extracting and analyzing. MinSup and MunConf are found from association rules of 15 users controlling. And the meaning of rule is found by user feedback.

Table 3. Extracted rules by confidence and support AND the number of user satisfied rules (user satisfied rules/total extracted rules)

	0.6%	0.7%	0.8%	0.9%
50%	125/985	123/341	118/178	70/72
60%	124/645	121/291	110/157	67/68
70%	119/548	113/194	108/135	67/68
80%	117/240	109/137	74/76	66/66
90%	117/155	98/104	68/68	65/65

In table 3 and 4, the association rules are the best meaningful when the number of user satisfied rules are

maximized. The association rules with usefulness for user are extracted from the best association rule among rules. But, when total extraction rules are so huge, these rules are very large for getting user feedback. For this, the rejection of user rules are considered. So the probability of user satisfied rules from total extracted rules are found. It may be that total satisfied rules are large when MinSup and MinConf are very small. But, according to the ratio of satisfaction, many total rules are extracted. Therefore, total user satisfied rule and ratio are considered by heuristic computing of support and confidence. In these results, when minimum support is 0.75% and minimum confidence is 80%, the best support and confidence are found. So the number of satisfied rules are 110 and the ratio of satisfaction is 98.2%.

Table 4. User satisfaction ratio of extracted rules

	0.6%	0.7%	0.8%	0.9%
50%	12.6%	36.0%	66.2%	97.2%
60%	18.9%	41.5%	70.0%	98.5%
70%	21.7%	58.2%	80.0%	98.5%
80%	48.7%	79.5%	97.3%	100%
90%	75.4%	94.2%	100%	100%

In following table, the learning rules are determined when user required minimum support is 0.75 and user required minimum confidence is 0.8. The results after learning make a comparison current and before.

Table 5. Change of extracted rules at each user
(# of satisfied association rules / # of extracted rules)

	First week	Second week	Third week
A	12/14	5/6	3/3
B	7/9	3/3	4/4
C	4/4	2/2	8/9
D	2/3	0/0	1/1
E	11/12	3/3	4/4
F	2/4	3/3	3/3
G	15/16	5/5	2/2
H	5/6	3/3	6/6

Table 5 shows that the number of rules are changed at each user and decreased considerably.

5 Conclusion

In this paper, the web mining technique for learning user feedback is proposed. and we extracted user interest from web log data. The fuzzy association rule was used for binary and quantitative association rules. This is used for user behavior using preprocessing web log file. The contents from useful page view are extracted using preprocessing with considered weights. Using fuzzy set with obscure idea, the user feedback for heuristic test in useful page was extracted. In our research, the fuzzy association rule by turning transaction file into fuzzy set was applied. This method is newer approach than existed one. The learning of input were added in user feedback. These have not been in existed methods. In these results, there are some advantages of user modeling approach. We propose a new approach in web usage mining using applying fuzzy logic to web mining. And this strategy was verified appropriate performance for extracting rules from web log data. In future works, fuzzy logic can be combined into machine learning algorithm. For example, the join fuzzy logic with support vector machine may be a good method for user modeling.

References

[1] R. Cooley, B. Mobasher, and J. Srivastava, "Web mining: Information and Pattern Discovery on the World Wide Web," In proceeding of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97) 1-3, 1997.

[2] R. Cooley, B. Mobasher, and J. Srivastava, "Data Preparation for Mining World Wide Web Browing Patterns," Journal of Knowledge and Information Systems, Vol.1,No. 1 8-19, 1999.

[3] A Gyenesei, "A Fuzzy Approach for Mining Quantitative Association Rules, TUCS Technical Reports No. 336, 2000.

[4] J. Han and M. Kamber, "Data Mining: Cincepts and Techniques," Morgan Kaufmann Publishers, 2001.

[5] C. M. Kuok and A. Fu, and M. H. Wong, " Mining Fuzzy Association Rules in Databases," SIGMOD Record 2-4, 1998.

[6] B, Mobasher, R. Cooley, and J. Srivastava, "Automatic Personalization Based on Web Usage Minng," Vol. 43 No. 8 Communication of the ACM 5-11, 2000.

[7] R. Srikant and R. Agrawal, "Mining quantitative association rules in Large Relational Tables," In Proceedings of the ACM SIGMOD Conference on Management of Data 3-8, 1996.

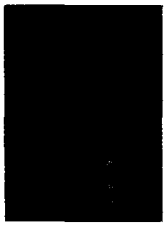
[8] S. M. Weiss and N. Indurkhya, "Predictive Data Mining," Morgan Kaufmann Publishers, Inc 51-79, 1998.

[9] H. J. Zimmermann, "Fuzzy Set Theory and Its Application," Kluewer Academic Publishers, 1996.

Jae-Sung Jang

Jae-Sung Jang received the M.S. degree in department of Computer Science from Sogang University, Korea, in 2000. He is currently with S/W department at LG Electronics Inc. His research interests include fuzzy logic and data mining.

Phone : +82-2-703-7626
Fax : +82-2-704-8273
E-mail : jaesung2@lge.com



Sung-Hae Jun

Sung-Hae Jun received the B.S., M.S., and Ph.D. degrees in department of Statistics from Inha University, Korea, in 1993, 1996, and 2001. He is currently with the artificial intelligence laboratory in the department of computer science at Sogang University, Seoul, Korea, where he is a Ph.D. candidate. His research interests include data mining, intelligent agents and cognitive modeling.

Phone : +82-2-703-7626
Fax : +82-2-704-8273
E-mail : shjun@ailab.sogang.ac.kr

Kyung-Whan Oh

Kyung-Whan Oh received the B.S. degree in mathematics from Sogang University, Seoul, Korea, in 1978, and the M.S. and Ph.D. degrees in computer science from Florida State University, Tallahassee, in 1985 and 1988, respectively. He is currently with the department of computer science at Sogang University, Seoul, Korea, where he is a Professor. His research and teaching interests include fuzzy system, cognitive science, knowledge discovery and data mining, intelligent agents and multi-agent systems, expert system and statistical learning.

Phone : +82-2-703-7626
Fax : +82-2-704-8273
E-mail : kwoh@ccs.sogang.ac.kr