Mining Generalized Fuzzy Association Rules from Web Logs

Rui Wu
School of Mathematics and Computer Science
Shanxi Normal University, Shanxi 041004, China
wurui19710905@126.com

Abstract

Discovery the association between web pages is an important task as the rapid growth of web data. This article uses the fuzzy method to discover generalized fuzzy association rules among the Web pages from Web logs. In the paper, whether a web page is visited or not and time duration on it are considered two important factors to reflect users' interest and preference. Numerical time duration is fuzzified into a corresponding fuzzy variable with membership values. The mined rule has the form as page(fuzzy duration) → page(fuzzy duration). These rules can reflect association among Web pages with fuzzy duration. By the analysis of the example, the generalized fuzzy association rules can be effectively mined in a sustainable computational period from Web user access patterns in Web logs.

Keyword: fuzzy web mining; fuzzy association rules; fuzzy variable; user access pattern

1. Introduction

WWW not only creates huge profits to web designers and operators, but also brings tremendous challenges. User can move from a web site to another one easily. If a web site does not meet the needs of users in a short time, then the chairman of the user relative to another site soon. Therefore, it is especially important to understand the needs and characteristics of Web users. Web mining can be regarded as extracting implicit structural information from structural or semi-structured data sets. In other words, it can be seen to discover data association from the given data sets. In the past many web usage mining algorithms were proposed to find interesting web access patterns from web logs[2, 8, 10]. Chen et al. [2] introduced the concept of using the maximal forward reference to break down user session into transactions for mining frequent traversal patterns. Spiliopoulou et al. [8] proposed an algorithm for building an aggregating tree from Web logs, then mining traversal patterns by MINT mining language. Xing et al.[10]didn't mine interesting web access patterns only by frequencies of visited web pages. They thought support and preference of web pages were used to extract interesting web access patterns. Deriving association rules from transaction databases is most commonly seen in data mining. It discovers relationships among items such that the presence of certain items in a transaction tends to imply the presence of certain other items. In the past, many researchers proposed several mining algorithms for finding association rules[1]. Recently, the fuzzy set theory has been used more and more frequently in intelligent systems because of its simplicity and similarity to human reasoning [3]. Hong et al. [3]and Lo et al. [6] used fuzzy set theory to efficiently discover relationships among items from database.

Whether a web page is visited or not discloses the interest of web users. Besides, duration on web pages is also an important factor to reflect users' interest. In addition, numerical or scalar data directly was mapped into a partition interval by traditional methods. This mapping approach is clearly not sufficient. Using fuzzy approach, an element for a collection doesn't only take 1 or 0 and is extended to [0,1]. This "soft" transition eases the border "hard" questions. Thus, this paper thus focuses on designing a sophisticated fuzzy web mining algorithm to find relationships among web pages. During the mining process, time durations are transformed as corresponding fuzzy linguistic variables. The proposal provided by Srikant and Agrawal is improved to discover interesting fuzzy association rules from web logs. Finally an example is given to verify the proposed method. Experiments show that the algorithm can effectively mine interesting rules for the understanding of the needs and characteristics of web users.

2. Preliminary

Zadeh [11] proposed the notion of fuzzy set in 1965. Later many researchers enriched and reinforced fuzzy theory, such as Nahmias [7] and Liu [4]. In this section, some basic concepts and results of fuzzy variable are reviewed.

978-1-4244-5934-6/10/$26.00 ©2010 IEEE

Definition 1 *A fuzzy variable ξ is defined as a function from a possibility space $(\Theta, \mathcal{P}(\Theta), \mathrm{Pos})$ to the set of real numbers, where Θ is a universe, $\mathcal{P}(\Theta)$ is the power set of Θ, and* Pos *is a possibility measure defined on $\mathcal{P}(\Theta)$.*

Definition 2 *Let ξ be a fuzzy variable on the possibility space $(\Theta, \mathcal{P}(\Theta), \mathrm{Pos})$. Then its membership function is derived from the possibility measure by*

$$\mu(x) = \mathrm{Pos}\{\theta \in \Theta | \xi(\theta) = x\}, \quad x \in R. \tag{1}$$

Definition 3 *(Liu and Liu [5]) The expected value of a fuzzy variable ξ is defined as*

$$E[\xi] = \int_0^\infty \mathrm{Cr}\{\xi \geq r\}\mathrm{d}r - \int_{-\infty}^0 \mathrm{Cr}\{\xi \leq r\}\mathrm{d}r, \tag{2}$$

provided that at least one of two integrals is finite. The credibility *of a fuzzy event $\{\xi \geq r\}$ can be represented by*

$$\mathrm{Cr}\{\xi \geq r\} = \frac{1}{2}[\mathrm{Pos}\{\xi \geq r\} + \mathrm{Nec}\{\xi \geq r\}]. \tag{3}$$

3. Algorithm of extraction fuzzy association rules

Web server access log file contains a large number of user's browsing information, which reveals the user's browsing behavior. Web log data original data needs preprocessing such as cleaning, user identification, session identification and transaction identification. Whether a web page is visited or not reflects the interest of web users. And time duration on a web page discloses the interesting degree of web users. Thus original web data is processed as {web page 1(time duration 1), web page 2(time duration 2),$\cdots$, web page n(time duration n)}, which is named as user access pattern. All user access patterns are stored in the database D_s.

Firstly, user access patterns are fuzzified. Time duration on each web page is denoted by a corresponding fuzzy linguistic variable with its membership value. The method can be described as follows.

Algorithm 1 Preprocessed algorithm

Input : n web access patterns, membership function sets

Output : n fuzzy web access patterns with fuzzy linguistic variable

step 1. Get the corresponding membership functions of time durations on all web pages by experts opinions or by methods introduced in [9]. Each membership function corresponds to a fuzzy region. Assume there exist s different fuzzy regions. Each fuzzy region is characterized as corresponding fuzzy linguistic variable $\lambda_i, (i = 1, \cdots, s)$.

step 2. Each time duration t_k on web page w_k visited by user D_i is transformed as fuzzy linguistic variable with its membership values.

$$u^1_{i,k}(w_k.\lambda_1) + \cdots + u^j_{i,k}(w_k.\lambda_j) + \cdots + u^s_{i,k}(w_k.\lambda_s)$$

The generalized mining algorithm combined with fuzzy variables and generalized data mining technology can be used to find web interesting rules from web access patterns. The algorithm is given as follows.

Algorithm 2 Fuzzy generalized mining algorithm

Input: n fuzzy web access patterns with fuzzy linguistic variable, predefined minimum support α,predefined minimum confidence β.

Output: fuzzy generalized association rules

step 1. Each web access pattern is transformed as corresponding fuzzy pattern with fuzzy linguistic variable using method introduced in algorithm 1.

step 2. Calculate the scalar cardinality $Count_j (1 \leq j \leq s)$ in every fuzzy region for each web access pattern $D_i (1 \leq i \leq n)$ by the following equation.

$$Count_j = \sum_{i=1}^{n} \mu^j_{k,i}, \tag{4}$$

where $\mu^j_{k,i}$ is the value of membership function of time duration t_i in the jth fuzzy region.

step 3. if $Count_{tk} = \max_{j=1}^{s} Count_{jk}$($t$ is the label of a fuzzy region), then the corresponding fuzzy linguistic variable of t is used to express the fuzzy characteristic of the web page.

step 4. Check the value $count_{tk} = \max_{j=1}^{s} count_{jk}$ in range $\lambda_j (1 \leq j \leq s)$. If it is greater than or equal to the predefined minimum support α, Put it into a frequent 1-itemset, that is $L_1 = \{\lambda_j \geq \alpha, 1 \leq j \leq s\}$.

step 5. Generate candidate 2-itemset C_2 from L_1.

step 6. For each item $S = (s_1, s_2)$ in the newly formed candidate 2-itemset C_2 do the following steps :

Compute the fuzzy value of each S by the following equation.

$$u_s = u^1_{i1,k1}(w_{k1}.\lambda_1) \wedge u^2_{i2,k2}(w_{k2}.\lambda_2). \tag{5}$$

If the minimum number of operations is for the intersection, then take the smallest $u_s = \min(u^{j1}_{i1,k1}, u^{j2}_{i2,k2})$. If it is greater than or equal to the

predefined minimum support α, Put it into a frequent 2-itemset L_2.

Calculate the scalar cardinality of the model S. If its scalar cardinality is higher or equal to the minimum support α, then the model S is put in the L_2.

step 7. If L_2 is empty, quit the algorithm. Otherwise continue the next step.

step 8. Set $r = 2$, where r is the number of items of the current frequent items.

step 9. Generate candidate item C_{r+1} from L_r by Apriori algorithm.

step 10. For each new $r + 1$-itemset C_{r+1} $(s_1, s_2, \cdots, s_{r+1})$ do the following steps:

[a.] Calculate the fuzzy value of S by the following equation.

$$u_s = u_1 \wedge u_2 \wedge \cdots \wedge u_{r+1} \tag{6}$$

where u_j is the membership value of the fuzzy item s_j. If the minimum operator uses intersection operation, then $u_s = \min_{j=1}^{r+1} u_j$.

[b.] Calculate the scalar cardinality of the model S by the following equation.

$$Count_s = \sum_{i=1}^{n} u_{is} \tag{7}$$

[c.] if $count_s$ is higher or equal to the predefined minimum support α, put the model S in L_{r+1}.

step 11. If L_{r+1} is empty, continue the next step. Otherwise set $r = r + 1$ and repeat the step 9-11.

step 12. For all frequent q-itemset L_q(including item $(s_1, s_2, \cdots, s_q)$, construct all possible association rules.

$(s_1 \wedge s_2 \wedge \cdots \wedge s_{k-1} \wedge s_{k+1} \wedge \cdots \wedge s_q)$.

step 13. Compute the confidence of the above association rules by the following equation.

$$Con = \frac{\sum_{i=1}^{n} u_{is}}{\sum_{i=1}^{n} (u_{is1} \wedge \cdots \wedge u_{isk}, u_{isk+1} \wedge \cdots \wedge u_{isq}} \tag{8}$$

where Con denotes the confidence of a association rule.

step 14. Retain the rules whose confidence is greater than or equal to the predefined minimum confidence β.

step 15. Output the gained rules to the users as the interesting rules.

4. Analysis of an example

In this section, an example is given to illustrate the proposed fuzzy generalized mining algorithm.

4.1 An example

Because browsing information of all users are stored in the Web log, so we extract user data from the Web logs. User data consists of two parts, in the Web log is expressed as (Url_{ik}, t_{ik}), in which Url_{ik} and t_{ik} denote the web page and the time duration on it visited by the ith user. Assume the extracted data is listed as follows.

Table 1. User access pattern from the log data

Client Id	Browsing sequences
1	(A,30), (B,42), (D,118), (E,91)
2	(A,92), (B,89), (F,120)
3	(A,50), (B,61), (D,42), (G,98), (H,115)
4	(A,70), (C,92), (G,85), (H,102)
5	(A,40), (B,35), (D,112)
6	(A,52), (B,89), (G,92), (H,108)

Assume membership functions of time duration on web pages by experts systems are shown as *short*(0,0,20,70), *middle*(20,70,90,120) and *long*(90,120,140,140).

According to the given algorithm, data mining process is described below.

Step 1: Fuzzy variables with membership values are adopted to describe users' data items in table 1. For example, the time duration on web page A 30 can be denoted as $(0.8(short), 0.2middle0.0long$. Web access pattern of the first user can be described as $(0.8(A.short), 0.2(A.middle))(0.56(B.short),$ $0.44(B.middle))(0.07(D.middle), 0.93(D.long))$ $(0.97(E.middle), 0.03(E.long))$.

Step 2: Calculate the scalar cardinality $Count$ in every fuzzy region for each web access pattern. As an example of fuzzy interval A.middle, its scalar cardinality=0.2+0.93+0.6+1+0.4+0.64=3.77. Repeat the step in other fuzzy intervals. Then scalar cardinality in each fuzzy interval on every web page can be gained.

Step 3-4: Determine count value in different fuzzy region of each web page whether greater than or equal predefined minimum support α. In the example, α is set 1.5. Each item higher or equal α is set in Ł$_1$.

Step 5-6: Generate C_2 from L_1. Get frequent 2-items from C_2.

Step 7-13: The possible association rules generated from frequent items are listed as follows:

If A=middle,then B=middle

If B=middle,then A=middle

If G=middle,then A=middle

If G=middle,then H=long

If H=long,then G=middle

Compute their confidence by the equation (7).

Step 14: Retain the rules whose confidence is greater than or equal to the predefined minimum confidence β. Assume minimum confidence $\beta = 0.8$, the following rules are retained.

If A=middle,then B=middle

If G=middle,then A=middle

If H=long,then G=middle.

Setp 15: Output the above rules to the users as the interesting rules.

4.2 Analysis of the results

The rule If A=middle,then B=middle is chosen as an example to analyze. Its confidence is 0.84. The rule could be interpreted as: If the access time for A page is middle, then duration on the page B is for middle. The confidence is 0.84. The rule is reflected not only the association between web page A and page B, but also web browsing on a time-dependent. The previous algorithms considered only the association between web pages. They neglected the page browsing time is also an important factor to reflect users' interest. In addition, fuzzy method is adopted to deal with numerical or scalar data. This "soft" transition eases the border "hard" questions.

5. Conclusions

As the rapid growth of data on the web, discovery and analysis of useful information from the web have become an increasingly important task. In the paper, an algorithm based on fuzzy theory is provided to mine fuzzy generalized association rules from user access patterns. Whether a web page is visited or not and time duration on it are considered two important factors to reflect users' interest. Therefore, the mined association rules not only are reflected the relationship between web pages, but also are related to the time duration on these web pages. Thus they reflect more exactly the user's interest and preference.

Acknowledgement

This work was supported by the National Natural Science Foundation of China No. 70802043 and the Shanxi Natural Science Foundation of China No.2008011029-2.

References

[1] R. Agrawal, T. Imielinksi, A. Swami, Database mining: a performance perspective, IEEE Transaction on Knowledge and Data Engineering 5 (6) (1993) 914-925.

[2] M. Chen, J. Park and P. Yu, Efficient data mining for path traversal patterns, IEEE Trans. Knowledge Data Engng, 10(1998) 209-221.

[3] T. Hong, C. Lee, Induction of fuzzy rules and membership functions from training examples, Fuzzy Sets and Systems, 84(1996) 33-47.

[4] B. Liu: Theory and practice of uncertain programming. Physica-Verlag, Heidelberg, 2002.

[5] B. Liu, Y. Liu: Expected value of fuzzy variable and fuzzy expected value models. IEEE Transactions on Fuzzy Systems. 10(2002) 445-450.

[6] W. Lo, T. Hong, S. Wang, A top-down fuzzy cross-level web-mining approach, in: The 2003 IEEE International Conference on Systems, Man and Cybernetics, 2003, pp.2684-2689.

[7] S. Nahmias: Fuzzy variable. Fuzzy Sets and Systems. 1(1978) 97-101.

[8] M. Spiliopoulou, "The Laborious Way from Data Mining to Web Mining", International Journal of Computer Systems Science and Engineering, 14(1999)113-126.

[9] X. Wang, M. Ha: Note on maxmin u/E estimation. Fuzzy Sets and Systems. 94(1998) 71-75.

[10] D. Xing, J. Shen, Efficient data mining for web navigation patterns, Information and Software Technology, 46(2004) 55-63.

[11] L. Zadeh: Fuzzy sets. Information and Control. 8(1965) 338-353.