Discovering Sequence Association Rules with User Access Transaction Grammars

Shi Wang Wen Gao Jintao Li
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080
{*shiwang, wgao, jtli*}*@ict.ac.cn*

Abstract

In this paper, We present a new approach to discover the associations between the access sequences, which is the sequence association rule discovery. In this approach, first we mine the Log in the web server to get the user access transactions, and then according to the regular grammar, we define the new user access transaction grammar in order to get the user sequence access transactions from the user access transactions. Subsequently we employ the association rule discovery algorithm to discover the sequence association rules. To evaluate this kind of rules, we propose the mutual information. The results of this approach can help the designer of the web site to understand the user access patterns better and according to the results the designer can adjust the structure of the web site.

1. Introduction

Because the World Wide Web is developing very quickly, the world of the Web generates volumes of data. So applying the approach of data mining to research these data, i.e. web mining, becomes a new promising important research field. The data in the world of the web mainly include the web pages, the web structures in web pages, and the server log. So the web mining includes web content mining, web structure mining, and web usage mining. When a user accesses a web site, he will leave his access log. A medium size site can record several megabytes per day. The log of a period of time is the good mining object. Through mining the Log, i.e. web usage mining, we can discover the user navigation patterns. This information can be used to improve the structure design of the web site.

Among the web pages, there is the rich structure information. The user navigation pattern has the close relation with this kind of the structure information. Through mining the server log, we can get the users' navigation patterns. The patterns can reflect the quality of the association among the web pages designed by the designer. The patterns also can be used to improve the structures of the web site and help users access the web site more conveniently.

The users' access patterns mainly include 1) the user navigation patterns [1], 2) the association rules among the web pages [2]. The former reflects the sequence discovery. The latter reflects the association degree among the discrete web pages. Because of the information of the net structures and the sequence characteristics in users' access we present a new approach to discover the sequence characteristics and the association characteristics, i.e. the sequence association rule discovery. This approach is based on the user access transaction grammar. The grammar is defined according to the regular grammar. Through the grammar, we can get all the access sequences in a user access transaction. Then the association rule discovery algorithm is employed to discover the sequence association rules, and the mutual information is used to give the rules better explanation. The sequence association rule can help the designer of the web site to understand the users' access patterns.

There are currently available some commercial log analysis tools [3]. However, these tools have limited analysis capabilities producing only results such as summary statistics and frequency counts of page visits.

Borges and Levene [1] apply the hypertext probabilistic grammar to discover the user navigation patterns and propose the use of entropy as an estimator of the statistical properties of the grammar. This approach can't be used to discover the association relation among different web page sets.

Cooley et al [2] firstly give the definition and classification of the web mining and give a system WEBMINER that mines the web usage. Their basic idea is to process the log in a web site, and organize the log data into the transactions. Then they can use the classical data mining approaches such as association rule discovery [4] to mine the data. Because they didn't consider the sequence access characteristics of the users, this approach can't discover the sequence association rules.

FootPrints [5] also takes an optimizing approach. Their idea is that visitors to a web site leave their "footprints" behind, over time, "paths" accumulate in the most heavily traveled areas, new visitors to the site can use these well

0-7695-0680-1/00 $10.00 © 2000 IEEE

worn paths as indicators of the most interesting pages to visit. WUM [6] improves this approach. It defines the g-sequences in order to mine the navigation patterns and gives a mining language MINT. These approaches only discover a kind of the local information; they can't be used to discover the association relations among different web page sets. Our approach's goal is to discover this kind of relations essentially.

The paper is structured as follows. Section 2 describes the mining objects. In section 3, we define the user access transaction grammar and how the grammar is used to generate the user sequence access transaction. We also give the algorithm of the generating process. In section 4, we define the sequence association rule and propose the mutual information to evaluate the discovered sequence association rule. In section 5, we compare our approach with Cooley's approach.

2. Mining objects

The Log is stored in the web server. Its format conforms to the standard of W3C [7]:

Before mining the user sequence association rules, we need transform the log to the user access transactions. L is the user's access log, its each entry $l \in L$ includes: the client user's IP address $l.ip$, the client user's identification $l.uid$, the accessed URL $l.url$, and the access time $l.time$. Then the access transaction t is:

Definition 1 the user access transaction:

$$t = < ip_t, uid_t, \{(l'_1.url, l'_1.time), \ldots, (l'_m.url, l'_m.time)\} >$$
$$where,\ for\ 1 \le k \le m, \quad (1)$$
$$l'_k \in L,\ l'_k.ip = ip_t,\ l'_k.uid = uid_t,\ l'_k.time - l'_{k-1}.time \le C$$

C is a stationary time window. Assuming there are n web pages in a web site, each page can be written: a_i, $i=1\ldots n$; then $A=\{a_1,\ldots,a_n\}$ represents the set of the web pages; the t can be written shortly: $t = <a_1', a_2', \ldots, a_m'>$. In the vector, $a_i'=l_i'.url$ and $a_i' \in A$. The algorithm that finds the access transaction is:

1.Pre-process the log. 2) Partition the log according to each user's IP address $l.ip$ to form each user's access set. 3) Partition each user's access set according to C to find access transactions. 4) Sort each visited web page in a transaction according to the visited time. 5) Sort all access transactions according to the visited time.

After processing the log, we get a set of the user access transactions.

3. User access transaction grammars

The goal of defining the user access transaction grammar is to get the order characteristics in a user access transaction. A web site node table is a finite nonempty set of symbol, such as $A=\{a_1,\ldots,a_n\}$. A^* denote the set of all finite user access sequences over A, including the empty sequence ε, it represent the possible path through which users access the web site. A^+ denotes the set $A^* - \{\varepsilon\}$. A sequence set L over A is any subset of A^*, a user access transaction grammar is a generative device capable of generating all the access sequence in a user access transaction.

Definition 2 User Access Transaction Grammar:

In a user access transaction t, the user access transaction grammar is 4-tuple $G=<V, \sum, S, P>$, where:

1.V is a set of the finite sequences: $V = \{S, A_1,\ldots, A_n\}$.

2.$\sum$ is the set of the visited web pages that are in transaction t: $\sum = \{a_1,\ldots,a_m\}$; $V \cap \sum = \phi$. The visited web page a_1 to a_m are sorted by visited time and given the corresponding subscript.

3.$S \in V$ is a unique start symbol.

4.P is a finite set of production rules with the general form $A_i \to a_i$ or $A_i \to a_i A_j$, where $A_i, A_j \in V$ and $a_i \in (\sum \cup \varepsilon)$, the subscript of each web page in the sequence A_j must be larger than t, and the subscript of the first web page in the sequence A_j must be equal to $t+1$.

In a user access transaction grammar G, a one-step derivation of sequence s_2 from sequence s_1, d: $s_1 \Rightarrow s_2$, occurs when a production is applied to s_1 to obtain s_2. A derivation, d: $s_1 \Rightarrow^* s_n$, is a finite sequence of one-step derivations that derives s_n from s_1. A sequence form is any derivation from the unique starting symbol S. The sequence set generated by a user access transaction grammar G is the set of all sequence forms composed only of web pages, $L(G) = \{s \in \sum^* | S \Rightarrow^* s\}$. A user access transaction grammar G is ambiguous if there is a sequence $s \in L(G)$, such that s has at least two distinct derivations from the starting symbol S. Otherwise, G in unambiguous.

A sequence form in a user access transaction grammar has at most one sequence; therefore, a production of type $A_i \to a$ is called a final production because it terminates the derivation process, and a production of type $A_i \to a_j A_j$ is called a transitive production. The length of a derivation, D, is the number of productions applied in the sequence derivation, which in a user access transaction grammar corresponds to the length of the generated sequence, i.e. the number of the web pages in the sequence. The sequence whose length is m is called m-sequence. The set of m-sequences is called m-sequence set.

Definition 3 the user sequence access transaction st: the st is the set of sequences generated from a user access transaction by a user access transaction grammar. The user sequence access transaction set, ST, is the set of the st. For example:

Table 1. The user access transactions

No	User access transaction
1	A_1, A_2, A_3
2	A_4, A_2, A_3, A_5
3	A_2, A_3, A_5
4	A_3, A_5, A_7

Table 2. The user sequence access transactions

No	1-sequence set	2-sequence set	3-sequence set	4-sequence set
1	A_1, A_2, A_3	A_1A_2, A_2A_3	$A_1A_2A_3$	
2	A_4, A_2, A_3, A_5	A_4A_2, A_2A_3, A_3A_5	$A_4A_2A_3, A_2A_3A_5$	$A_4A_2A_3A_5$
3	A_2, A_3, A_5	A_2A_3, A_3A_5	$A_2A_3A_5$	
4	A_3, A_5, A_7	A_3A_5, A_5A_7	$A_3A_5A_7$	

The algorithm that generates user sequence access transaction from the user access transactions is:

Algorithm: GUSAT

Input: $t=< t_1,\ldots,t_m >$

Begin:

$k := 1$;

$S^k := \{t_1,\ldots,t_m\}$; /* S^k is the set of the k-sequences */

While $k \le m$

For each $s \in S^k$

$p :=$ position(t,s); /* In string t, getting the start position of the string s */

If $(p+k+1) \le m$ then

$s :=$ merge(s, t_{p+k+1});/* Adding t_{p+k+1} to the tail of the s*/

$S^{k+1} := S^{k+1} \cup \{s\}$;

End If;

End For;

$k := k + 1$;

End While;

End.

Output: S^k, $k = 1, \ldots, m$

Applying this algorithm to each user access transaction will generate its user sequence access transaction. Finally we get the set of the user sequence transactions.

4. Sequence association rule discovery

The association rule discovery [4] is mainly used to discover the relations between the set of the large items. To discover the sequence association rules, we give these below definitions:

Definition 4 the support of the sequence s, *support*(s): Given a sequence s, in *ST*, *support*(s) is the number of the transactions that contain the sequence s.

Definition 5 the sequence association rules:

The s and s' represent two different sequences in *ST*. Their *supports* are bigger than a given *support* threshold.

Then their *confidence* is:

$$confidence(s,s') = \frac{support(s,s')}{support(s)} \quad (2)$$

If $confidence(s, s') \ge \theta$, θ is a given threshold of *confidence* (such as 5%), $s \rightarrow s'$ constitutes a sequence association rule. Note that the *support*(s, s') in definition 5 represents the number of transactions that contain s and s' together in *ST*.

We use the AprioriHybrid [4] algorithm to discover the sequence association rule.

To evaluate the sequence association rule, we propose the mutual information [8], that is, if two sequence frequently together, they probably have high association:

$$MI(s,s') = \log\frac{P(s,s')}{P(s)P(s')} \quad (3)$$

Because this metric does not consider the effect of the absence of either or both sequence in a user sequence access transaction, we proposed the expected (or average) mutual information [9]. Note that if two sequences always appear together or not at all, they more likely have high association, it captures the effects of sequence absences:

$$EMI(s,s') = \sum_{s,\bar{s}\in S}\sum_{s',\bar{s}'\in B} P(S,S')\log\frac{P(S,S')}{P(S)P(S')} \quad (4)$$

Through computing the mutual information and the expected mutual information to the discovered sequence association rule, more valuable information can be discovered in the set of the user sequence access transactions that are generate by the user sequence access transaction grammar.

Comparing with Cooley's approach [2], they process the user access transaction and don't consider the correlation in the visited web pages in a user access transaction. The association rule set generated by our approach includes that generated by Cooley's approach. In our approach, the correlation generated by the user sequence access grammar can be used to improve the design of the web site. For example, in Figure 2, the discovered sequence association rule can be used to adjust the structure of the web site.

Through mining the *ST*, if in the set of 1-sequence we discover $(B,A)\rightarrow(D,C)$, its *confidence* is 80% (this rule is equal to the result of Cooley's approach). This rule is explained that 80% users that accessed B, A also accessed D, C. If through mining the 2-sequence set, we can get a sequence association rule: $AB\rightarrow CD$, its *confidence* is 70%. It is explained that 70% users that orderly accessed A and B also orderly accessed C and D. So we can add a hyperlink from B to C in order to facilitate the users' access. In other words, through mining the sequence association rules, we can understand the discovered knowledge better and these rules can be used to improve the structure design in web site.

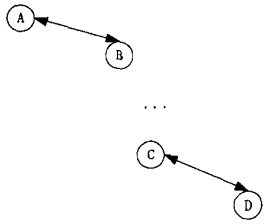

Figure 1. There isn't direct hyperlink between *A,B* and *C,D*

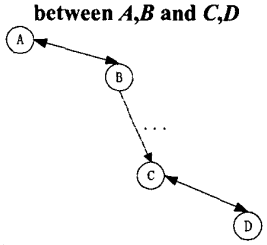

Figure 2. Adding the hyperlink between *B* and *C*

5. Experiments

We choose the Log in the WWW server of Institute of Computing Technology, Chinese Academy of Sciences (http://www.ict.ac.cn). The experiment data include one year's access data from 1998.11 to 1999.11. The web site includes 352 html web pages. The Log is 147M. It includes 174,9934 entries. The algorithm that finds the access transaction finds 10399 user access transactions. The average length of each transaction is 8.8. In the experiment, the hardware is a PC with PentiumIII450, 64M RAM, and 6G hard disk. The software is Windows NT operation system.

1. Comparing with Cooley's approach:

Discovering the sequence association rules in 1-sequence set, which is equal to Cooley's approach, we can get association rules:

Table 3. The 1-sequence association rule

(1-sequence) to (1-sequence), *support* ≥10,*confidence* = 25%
(/cjc/cjccw.html,/cjc/cjcc.html,/cjc/introc.html, /cjc/cjcw2.html) → (/cjc/contc.html,/cjc/abstc.html)

According to our approach, relative to the 1-sequence association rule, the discovery result is:

Table 4. Some discovered sequence association rules

The sequence association rule, *support* ≥5,*confidence* ≥ 1%	*confidence*
/cjc/cjccw.html → /cjc/cjcc.html /cjc/introc.html /cjc/cjcw2.html	98%
/cjc/cjccw.html /cjc/cjcc.html /cjc/introc.html /cjc/cjcw2.html → /cjc/contc.html /cjc/abstc.html	15%
/cjc/cjccw.html /cjc/cjcc.html /cjc/cjcw2.html /cjc/introc.html → /cjc/contc.html /cjc/abstc.html	7%
/cjc/cjccw.html /cjc/cjcc.html /cjc/introc.html /cjc/cjcw2.html → /cjc/abstc.html /cjc/contc.html	2%
/cjc/cjccw.html /cjc/cjcc.html /cjc/cjcw2.html /cjc/introc.html → /cjc/abstc.html /cjc/contc.html	1%

Apparently, to the discovered rules, our approach gives better explanation than Cooley's approach.

2.The mutual information:

The mutual information can give the better explanation to the sequence association rules than the traditional discovery metric method (*support*(*s*)=2047):

Table 5. The associations between different sequences

No	Sequence association rule $s \rightarrow s'$	*support* (s')	*support* (s, s')	*confidence*	*MI*	*EMI*
1	/cjc/cjccw.html → /cjc/contc.html /cjc/cont98c.html	25	07	.2%	0.6383	.0054
2	/cjc/cjccw.html → /cjc/absc.html /cjc/abstc.html	12	81	3.7	.6604	.0158
3	/cjc/cjccw.html → /cjc/othersc.html /cjc/ly.html	07	09	5%	.5862	.0132
4	/cjc/cjccw.html → /cjc/abstc.html /cjc/contc.html	50	33	6%	.6922	.0357

Comparing the formula 2 and the formula 3, after adding the $P(s')$, the formula 3 can explain the correlation between s and s', i.e. if $P(s)$ and $P(ss')$ aren't changed and if $P(s')$ increases, the mutual information of the two sequences will reduce. The bigger $P(s')$ will reduce the mutual information. The mutual information explains thej discovered rule better. For example, in Table 5, about the second rule and the third rule, the *confidence* of the second rule is lower than that of the third rule, but it has the higher mutual information and expected mutual information than the third rule, which shows that the path: /cjc/cjccw.html /cjc/absc.html /cjc/abstc.html will be a users' access path more possibly than the path: /cjc/cjccw.html /cjc/othersc.html /cjc/ly.html.

3. The performance of the algorithm:

We employ the AprioriHybrid algorithm to get the sequence association rule. The Figure 3 gives the execution time of each pass. Figure 4 gives the relations

between the minimum *support* and the execution time:

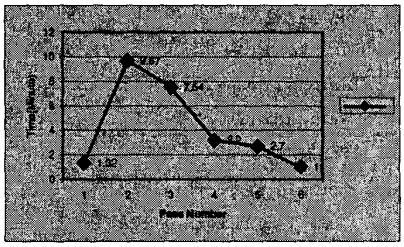

Figure 3. The execution time of each pass (The minimum *support* is 3)

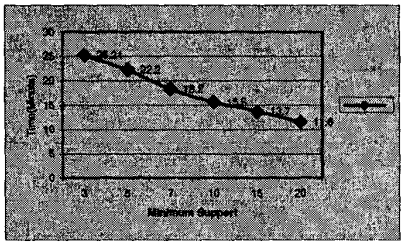

Figure 4. The relations between the minimum *support* and the execution time

6. Conclusions and future work

In web mining, applying association rule discovery can discover the associations between different web pages accessed by users. Because there is the rich structure information in the web site and the access of the users must conform to some kinds of access sequences, in this paper, we can present a new approach to discover the associations between the access sequences, which is the sequence association rule discovery. This approach includes the Cooley's approach, and it discovers the associations between different user access sequences. In this approach, first we mine the Log in the web server to get the user access transaction, and then according to the regular grammars, we define a new user access transaction grammar in order to get the user sequence access transaction from the user access transactions. Subsequently we employ the association rule discovery algorithm to discover the sequence association rules. To evaluate this kind of rule, we propose the mutual information. The result of this approach can help the designer of the web site to understand the user access pattern better and according to this result the designer can adjust the structure of the web site.

There are some characteristics in our approach: 1) It discovers the sequence association rule. 2) It mines periodically and offline. 3) The mining object is the interactive action and the common interest; the mining result faces up to all users. 4) It doesn't require the information of the certain user or user class. 5) The approach can discover the rules in different web classification sets in web site, it is not a local approach, and the discovered sequences possibly haven't direct hyperlinks.

According to this approach, the designer of the web site can discover the association relation in different web page sets; he also can add the direct hyperlinks to facilitate the users' access. The future work will apply this approach to predict the user's current access actions, and recommend the personalized next web pages in real time.

References

[1] J. Borges, and M. Levene, "Data mining of user navigation patterns", In *proc. of the Web Usage Analysis and User Profiling Workshop*, San Diego, California, August 1999, pp. 31-36.

[2] R. Cooley, B. Mobasher, and et al, "Data Preparation for Mining World Wide Web Browsing Patterns", *Knowledge and Information Systems*, 1999, 1(1).

[3] R. Stort, "Web Site Stats: tracking hits and analyzing traffic", Osborne McGraw-Hill, 1997.

[4] R. Agrawal, and Srikant, R. "Fast algorithms for mining association rules", *In Proc. Of the 20th VLDB Conference*, Santiago, Chile, 1994, pp. 487-499.

[5] A. Wexelblat, and P. Maes, "Footprints: History-rich web browsing", In *Proc. Conf. Computer-Assisted Information Retrieval(RIAO)*, 1997, pp. 75-84.

[6] M. Spiliopoulou, "The laborious way from data mining to web mining", *Int. Journal of Comp. Sys., Sci. & Eng.*, Special Issue on "Semantics of the Web", Mar.1999.

[7] A. Luotonen, "The common log file format", http://www.w3.org/pub/WWW/, 1995.

[8] L. Chen and K. Sycara, "Webmate: A personal agent for browsing and searching", In *Proc. 2nd Intl. Conf. Autonomous Agents*, 1998, pp. 132-139.

[9] R. Rosenfeld. "A maximum entropy approach to adaptive statistical language modeling", *Computer, Speech, and Language*, October 1996.