J. of university of anbar for pure science : Vol.4:NO.2 : 2010

ISSN: 1991-8941

Web Usage Pattern Discovery Using Time Stamp Association Rules

Nidal Ali Hussein

University of Baghdad – College of Islamic Science

Received: 12/5/2008 Accepted: 2/12/2009

Abstract: Association rules are typically used to describe what items are frequently bought together. One could also use them in web usage mining to describe the pages that are often visited together .The goal of web usage mining is to extract useful knowledge from the data that web servers store about the behaviour of the customers. In this paper, we introduce an extension to association rules by the introduction of time stamp that can give us a better insight into the data. Subsequently, the introduced concepts are used in an experiment to pre-process log files for web usage mining. We also describe how the method could be useful for market basket analysis and give an overview of related research. The paper is concluded by some suggestions for future research.

Key words: Web , Pattern , Discovery , Stamp , Association Rules

Introduction

In[1], the problem of mining association rules was first outlined. The concept was applied on the data of a supermarket. An example of an association rule in that context is "90% of the transactions that contain bread and butter also contain milk." Afterwards the approach of association rules was also adopted for the analysis of web site traffic. The resulting association rules indicate which pages are often requested together. With this information it is possible to forecast the next pages a visitor will frequent [7] .

A disadvantage of association rules however is that they don't take into account the sequential information that is available in some data. For example, it is interesting to know that pages A and B are often visited together but it might be even more interesting to know that page B is always visited immediately after page A. In this paper, we propose an extension to association rules that takes the timing information of the data into account.

This paper is organized as follows. In the next section a brief introduction to association rules is given. Afterwards, we extend the concept of association rules to take into account time stamp. In the third section of this paper, a practical application of the proposed extension is discussed. Finally, we conclude by some suggestions for future research.

Association Rules

Let D be a database of transactions. Each transaction consists of a transaction identifier and a set of items {i1, i2, …, in} selected from the universe I of all possible descriptive items. In Table 1 an example database is shown containing four transactions.

Table 1: DataBase with 4 Transactions.

Transaction	Items
1	A, B, C
2	A, C
3	A, D
4	B, E, F

What the items represent depends on the application. For example, the items could be the different products bought by a customer (or, as in this paper, the web pages that someone visited). An association rule is an expression of the form: $X \Rightarrow Y$ where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$. Each association rule is characterized by means of its support and its confidence defined as follows:

$$\sup(X \Rightarrow Y) = \frac{\text{number of transactions containing } X \cup Y}{\text{total number of transactions}}$$

$$conf(X \Rightarrow Y) = \frac{\sup(X \Rightarrow Y)}{\sup(X)}$$

For the example database, it can be verified that the rule A ⇒ C has support 50% and confidence 66.7%. According to the above definitions, the support measure can be considered as the percentage of database transactions for which (X ∪Y) evaluates to true. Equivalently, the confidence measure is understood to be the conditional probability of the consequent given the antecedent. Association rule mining essentially boils down to discovering all association rules having support and confidence above user-specified thresholds, minsup and minconf, for respectively the support and the confidence of the rules [2] .

In this paper, we propose an extension to the association rule framework by introducing time stamp. Suppose we have the following database D which is similar to the one in Table 1 but with every item is a time stamp associated. As only the timing within a transaction is considered of importance, the time stamp of the first item in every transaction is set equal to 0.

Now consider the following definitions where X

example, from the 100% confidence of the rule G ⇒ A we can conclude that page G is always visited in combination with page A. As the forward confidence of this rule is 0% we can deduct that page A is always visited before page G. Combining these two measures gives us a better insight into the data. It is possible to define backward support and confidence in a similar way:

***Table 2*: DataBase Containing 7 Transactions.**

Transaction	Items	Time Stamp
1	A B C	0 4 5
2	B A	0 100
3	A B D G	0 4 4 67
4	A B G	0 2 4
5	A D	0 6
6	B D	0 8
7	A B G	0 5 10

$$backward\ \sup(X \Rightarrow Y) = \frac{\text{number of trans. containing X} \cup \text{Y whereby time}(X) > \text{time(Y)}}{\text{total number of transactions}}$$

$$backwardco\quad nf\ (X \Rightarrow Y) = \frac{backward\quad \sup(X \Rightarrow Y)}{\sup(X)}$$

$$\text{timesupport}(t_1, t_2)(X \Rightarrow Y) = \frac{\text{number of trans. containing X} \cup \text{Y whereby } t_1 \le \text{time(Y) - time(X)} < t_2}{\text{total number of transactions}}$$

$$\text{timeconf}(t_1, t_2)(X \Rightarrow Y) = \frac{\text{timesupport}(t_1, t_2)(X \Rightarrow Y)}{\sup(X)}$$

We can generalize these definitions to:

$$forwardconf\ (X \Rightarrow Y) = \frac{forward\ \sup(X \Rightarrow Y)}{\sup(X)}$$

,Y ∈ I and X∩Y = Ø.

For the example database it can be verified that the forward support of rule A ⇒ B corresponds to 4/7=57% and the forward confidence equals 66.7%. In contrast with the calculation of the normal support and confidence, the second transaction is not counted because B proceeds A. By using the timing information that is available in the database it is possible to give more meaning to the association rules. For

whereby t1 and t2 are integers and t1 ≤ t2. We can see that these definitions are generalisations of the association rule framework.

∀ti, tj, tk: if ti ≤ tj then

timesup(tj, tk)(X⇒Y) ≤ timesup(ti, tk)(X⇒Y) ≤ sup(X⇒Y)

timesup(tk, ti)(X⇒Y) ≤ timesup(tk, tj)(X⇒Y) ≤ sup(X⇒Y)

timeconf(tj, tk)(X⇒Y) ≤ timeconf(ti, tk)(X⇒Y) ≤ conf(X⇒Y)
timeconf(tk, ti)(X⇒Y) ≤ timeconf(tk, tj)(X⇒Y) ≤ conf(X⇒Y)
For t1=-∞ and t2=+∞ the foregoing definitions correspond to the normal definitions of support and confidence. For t1=0 and t2=+∞ (t1=-∞ and t2=0) these definitions correspond to the definitions of forward (backward) support and confidence.

Properties

From the definitions of timesupport and timeconfidence we can easily derive two associated measures, which we call support ratio and confidence ratio.

$$\text{support ratio}(t_1, t_2)(X \Rightarrow Y) = \frac{\text{timesup}(t_1, t_2)(X \Rightarrow Y)}{\sup(X \Rightarrow Y)}$$

$$\text{confidence ratio}(t_1, t_2)(X \Rightarrow Y) = \frac{\text{timeconf}(t_1, t_2)(X \Rightarrow Y)}{\text{conf}(X \Rightarrow Y)}$$

As the timesupport (timeconfidence) of a rule is always smaller or equal than normal support (confidence), it is easily derived that each ratio lies in the range of 0 to 1. Furthermore, for a given association rule, both ratios are equal and express the conditional probability of X and Y occurring in the time window defined by t1 and t2 given the fact that X and Y appear in the transaction.

In the example database, every item appears only once in every transaction. In many real-life applications this is obviously not the case. For example, in web usage mining, a user can visit some pages more than once. In that case, there are multiple ways to calculate the time difference between two items [6]. Consider the transaction in Figure 1. In this transaction, items A and B appear several times. To calculate timesup(t1,t2)(A ⇒B) we need to know if t1 ≤ time(B)-time(A) ≤ t2 is valid for this transaction. If some items appear more than once, the time difference can be defined in many ways. It is possible to count a transaction only if t1 ≤ time(B) - time(A) ≤ t2 is valid for all occurrences of A and B. A second approach could be to count a transaction only if the equation is valid for at least one occurrence of A and B. A third approach could consist of calculating the average time difference between the items. If this average time difference falls between the limits imposed by t1 and t2 the transaction is counted. Many other approaches that give different results could be proposed. For our research, we subtracted the time stamps of the first occurrences of A and B. For the example transaction this results in a time difference of 4.

In the next section we discuss a practical application of the proposed extension to the association rules framework.

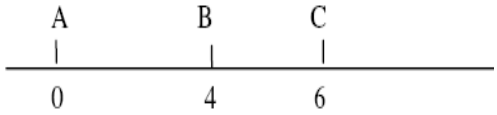

Figure 1: Transaction with Multiple Items.

Practical Application

Web usage mining is defined as the application of data mining techniques to discover usage patterns from web data. In other words, the goal of web usage mining is to extract useful knowledge from the data that webservers store about the behaviour of the customers. It is generally agreed that web usage mining consists of three main phases: preprocessing, pattern discovery and pattern analysis [8] .

In this section, the focus lies on the first phase: preprocessing. During the preprocessing phase all the necessary operations to transform the data in a form suited for the chosen type of analysis are performed. As we have opted to mine for association rules, we will transform our raw logfiles in a form similar to Table 1. In[5], a survey is given of the different operations one must perform in the preprocessing phase. We only consider the first step of the preprocessing phase, namely data cleaning. Raw logfiles contain a lot of lines that are irrelevant for web usage mining. These lines must be deleted before applying the mining techniques. The principle hereby is that with every action of the visitor usually a click of the mouse should correspond one line in the logfiles. When a visitor requests a page, this request will be logged, but it is not the only request that will appear in the logfiles. The HTML-code of the page indicates which pictures the browser should show. When the browser analyzes the HTML-code, it will send requests for these pictures. So if there are four pictures on the requested page, there will appear five lines in the logfiles: one for the HTML-page and one for every picture. These requests for pictures must be deleted from the logfiles because the user did not explicitly ask for them. Cleaning the logfiles from pictures and photos is quite easy.

It suffices to examine the extensions of the requested files and extensions that correspond with pictures, such as .jpg and .gif, should be deleted. For a similar reason requests for directories and stylesheets are deleted [3] .

All of the preceding steps are independent of the analyzed logfile and can be performed by a standard program. However, sometimes there are

additional steps required that are specific for the analyzed data. For example, if a visitor types the address of the home page of the site we studied, he is transferred several times to other pages before he finally sees the real homepage. These transfers happen automatically and are almost imperceptible for the users, but they can be seen in the logfiles. Only one of these requests should be kept in the logfiles because the user performed only one action [6]

It was also noticed that many requests occurred for the page http://www.oursite.com/showit.htm. This page was used as a script to retrieve images from a database. Requests for this page were embedded in the HTML-code of other pages. Because users did not explicitly ask for showit.htm, all requests for this page were eliminated. We found several other ways of how a user could generate multiple lines in the logs with only one click of the mouse. We show how the proposed extension of association rules is capable of detecting these cases. The removal of these double requests results in fewer but more interesting association rules as the evident ones are left out. This facilitates considerably the analysis of the rules.

For the analysis, we used one week of logdata from the work of PhD student (Hayder Mahmood Salman), he was take the logdata from the server of Al-Sulaimania University/Computer Center by using packet sniffing technique. Requests for pictures were deleted from these data. Afterwards the heuristic proposed by [3] was used to identify the different visitors. The requests from a visitor were divided into multiple sessions whenever a time-out of more than 30 minutes occurred between two successive requests [4]. After performing these operations, the data were in a form similar to Table 2. On these data a slightly modified version of the Apriori-algorithm was executed several times with different values for t1 and t2. For minsup and minconf, we opted for respectively 2% and 30%. This resulted in the detection of 208 rules with one item in both antecedent and consequent. For each of these rules it was checked manually if both the pages from head and body of the rule could be requested with one single action of the visitors. In the rest of this paper we will call these rules artificial. An example from such an artificial rule is: http://islampaik.org/⇒ http://www.msn.com/, i.e. when the http://www.msn.com/ page is accessed in a session, the http://islampaik.org/ page is also accessed. From the 208 discovered rules, 52 were found to be artificial. For t1=-25 and t2=25 the confidence ratio of all the rules was calculated. An overview is given in Table 3 and Figure 2.

Table 3: Overview.

Conf. ratio(-25,25)	Rules	Artificial rules	Non-artificial rules
>0,6	39	39	0
0,5-0,6	8	6	2
0,4-0,5	6	1	5
0,3-0,4	8	2	6
0,2-0,3	14	4	10
<0,2	133	0	133
total	208	52	156

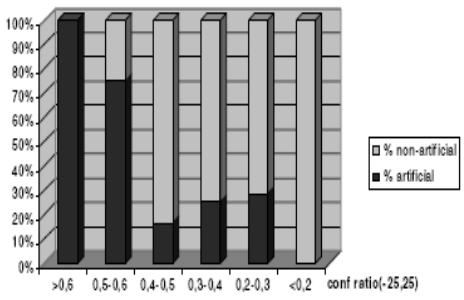

***Figure 2*: Overview.**

All of the rules with a confidence ratio of more than 60% are artificial. Rules with a lower confidence ratio on the other hand are mostly nonartificial. By using 50% as a cut-off only 9 rules are incorrectly classified. Hence, the confidence ratio can be used to separate the artificial rules from the others. Mild variations of t1 and t2 made little difference in the outcome of the experiment.

A drawback of the above method is that it relies on the response time of the server. Most of the misclassifications in our experiment occurred

because it took the server a long time to generate an automatically requested page. In that case, the confidence ratio will become small and a misclassification is likely to occur. The opposite case was also observed. Pages with only little text and one possible link.

We have shown that the proposed extension of association rules is capable of detecting the rules that are generated automatically. Because these rules are not affected by the browsing behaviour of the visitors, they are of no value for the business. Moreover, pages that are requested automatically could be deleted from the logs. Their removal results in less but more interesting association rules. This facilitates considerably the burden of post-processing.

The proposed extension could also be used for recommendation purposes. Rules with a high forward confidence indicate possible future requests. However, to be useful for recommendation purposes, it is needed to expand the proposed concept by allowing more than one item in the antecedent of the rules.

Market basket analysis was the first application field of association rules. There was no ordering of the items within a transaction because all the items bought were placed in a basket and then finally paid. It was impossible to detect the order in which the customer took the products from the shelves. Nowadays, a large number of supermarkets are doing experiments with self scanning devices. The customers scan the code of every product they buy and when finished shopping they go with the scanner and products to the counter. At the counter the payment is done and no clerks are needed to scan all the products again. However, some random checks are performed to prevent fraud from the customers. This method not only reduces the amount of employees needed or the waiting times but also creates a lot of worthwhile information. The marketeers can see how customers wander through the shop and adopt the layout of the shelves to improve sales. It can be expected that ordinary association rules will not suffice to analyse this kind of sequential data. It is for example impossible to detect with association rules if customers go mostly from shelf A to shelf B or vice versa. It can not be seen why diapers and beer are frequently sold together. Were diapers the first objective and was the beer only picked up when passing by [1]? The proposed extension might prove useful for answering some of these questions.

Conclusion

In this paper, we presented an extension of the basic association rule framework that is capable of analyzing sequential data .It is interesting to know that pages A and B are often visited together but it might be even more interesting to know that page B is always visited immediately after page A. In this paper , we propose an extension to association rule that takes the liming information of the data into account. The concept was tested on the log data of Computer Centre / Al-Sulaimania University and it was shown how it helped filtering out the interesting rules. It was also suggested to use the proposed extension for the analysis of data generated by self-scanning devices.

In future research, the introduced concept will be used to analyze the logs from other web sites and it will be expanded to allow more than one item in the body of the rules.

References

[1] Agrawal R., Imielinski T. & Swami A., 1993. Mining Association Rules between Sets of Items in Massive Databases. In Proceedings of the ACM GMOD International Conference on Management of Data, Washington D.C., USA. pp. 207-216.

[2] Agrawal, R., & Srikant, R., 1995. Mining sequential patterns. In Proceedings of the 11th International Conference on Data Engineering. pp. 3-14.

[3] Ahmed Tariq Sadiq and Hayder Mahmood Salman, 2008. Packet Sniffing to Prepare Data for Web Usage Mining. Iraqi Journal of Information Technology, No. 2.

[4] Catledge L. & Pitkow J., 1995. Characterizing Browsing Strategies in the World-Wide Web. Journal of Computer Networks and ISDN systems, Volume 27, nr. 6. pp. 1065-1073.

[5] Cooley R., Mobasher R. & Srivastava J., 1999. Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information Systems, volume 1. pp. 5-32.

[6] Cooley R., 2000. Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data. PhD. Thesis. University of Minnesota.

[7] Mobasher B., Dai H., Luo T. & Nakagawa M., 2002. Using Sequential and Non-Sequential Patterns for Predictive Web Usage Mining Tasks. In Proceedings of the IEEE International Conference on Data Mining. pp. 669-672.

[8] Srivastava J., Cooley R., Deshpande M., Tan P-N., 2000. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. Web Data, SIGKDD Explorations, Vol. 1, Issue 2. pp. 12-23.

استكشاف أنماط استخدام شبكة الانترنيت باستخدام قواعد الارتباط المحدد بالوقت

نضال علي حسين

E.mail: scianb@yahoo.com

الخلاصة :- إن قواعد الارتباط تستخدم لوصف أي المواد التي تشترى بشكل دائم مرتبطة مع بعضها البعض (أكثر من مادة مرتبطة بالشراء معاً) ، وتستخدم أيضاً لوصف صفحات البحث في شبكات الانترنيت التي غالباً ما تزار من قبل أكثر من زبون (مستخدم) .إن الهدف من استخدام البحث في شبكات الانترنيت هو الاستخراج المعرفة المفيدة من البيانات التي تم خزنها حول سلوك الزبائن (أي المستخدمين) .نقترح في هذا البحث تطوير طريقة قواعد الارتباط وذلك باستخدام محددات الوقت ، التي ممكن أن تعطينا طريقة أفضل في البحث عن المعلومات ومعالجة سجلات البيانات على شبكات الانترنيت كتحليل أسعار السوق وغيرها .ويقدم هذا البحث أيضاً استنتاجات واقتراحات لبحوث مستقبلية ذات العلاقة .