VISITOR NAVIGATION PATTERN PREDICTION USING MARKOV MODELS, ASSOCIATION RULES AND AMBIGUOUS RULES

[1]EiTheint Theint Thu, [2]Khaing Min Kyu, [3]Hlaing Htake Khaung Tin
University of Computer Studies, Hinthada, Myanmar[1], University of Computer Studies, Magway, Myanmar[2], University of Computer Studies, Hinthada, Myanmar[3]
eitheinttheintthu@gmail.com[1], kminkyu84@gmail.com[2], hlainghtakekhaungtin@gmail.com[3]

ABSTRACT

There are large number of Web sites which consist of many web pages. It is more difficult for the users to quickly get their target pages. The main aim of this paper is to only implement Visitor Navigation Pattern Prediction Using Markov models, Association Rules and Ambiguous Rules. The paper uses traveling paths that assist visitors to navigate the visiting places based on the past visitor's behavior. In this article, Markov models and association rules and ambiguous rules were used to resolve ambiguous web access predictions .An improved method that organizes Markov models, association rules and ambiguous rules and combines web pages into a web site for prediction. This method can offer better predictions than using each method alone and other traditional models. It uses web usage mining techniques for recommending a visitor which (next)paths is closely the most popular paths in Myanmar.

INTRODUCTION

Myanmar is one of the most interesting countries to explore in Asia. The Travels and Tours is a Myanmar based travel agency website. In other words, travelling to Myanmar is very similar to traveling back and forth to other geographical areas, not just one trip. The large number of cities on many visiting places in Myanmar has raised. In the website, visitors often have the first navigational question such as where can I go? The website owner can advise you the popular paths in there. The system refers to tracking the visitor's past behavior and gathering increasing amounts of visitor's information in web log file, it can predict the next visiting places [1].

Internet mining is defined as the application of data mining techniques to discover custom web browsing patterns from web log data, in order to understand and better meet the needs of web applications. The job presented in this document is to search a website log file for information about a website and its users, and use this knowledge to help users navigate and find a recommended website effectively and efficiently.

One of the major data sources for this study is web server log data, which tracks user activity while browsing the web. Recently, it has become known and discussed to predict the behavior of Internet users and their next move. The outcome of precise predictions can be consumed to recommend products to customers, propose useful links, and pre-submit, pre-read, and cache web pages to reduce access latency [2]. There are different forecasting methods, but the most popular are Markov models, association rules and ambiguous rules. Markov models are exercised to determine the next page that a website user should go to, based on the sequence of earlier visited pages. Association rules can be consumed to determine the likely next requests for a web page based on statistically significant correlations. To determine with the same levels of confidence, ambiguous rules are used. In Markov models, the goal is to create prediction models to guess which web pages might be requested next time and Transition Probability Matrix (TPM) is created and the expectations for web sessions are simple. Lower-order Markov models lack precision due to the limited amount of navigation history, while higher-order Markov models tend to result in higher spatial complexity. On the other hand, association rules and ambiguous rules pose the problem of determining the correct prediction from a set of rules. This paper uses combination of Markov models, association rules and ambiguous rules to provide better forecast accuracy. The

consequence is that only the highest probability of the condition is taken into account. Therefore, the prediction rule with the highest probability of the condition is selected.

RELATED WORK

Recommendation systems are one of the first web forecasting applications. There are many applications to analyze web browsing user behavior in web crawling analysis. Analyzing user behavior while browsing the Internet can help improve the organization of websites. Web Personalization and Responsive websites are some of the frequently used applications in internet usage analysis (Internet usage extraction). The most commonly used approach is web crawling, which contains many models such as Markov models, association rules and clustering. The Markov model is a popular approach for predicting which pages to go next.

Markov models can be used to determine the next state based on the previous state but Lower order Markov models do not take the history into account in detail and the precision is very low while the High order Markov models have high complexity. Three approaches have been widely used to solve this problem, for instance frequency-pruning, error-pruning and support-pruning to reduce state space complexity. Also, the web page prediction techniques are applied by combining with K-means clustering and Latest Substring Association (LSA) rule mining method. The prediction can perform on the clustering sets rather than the real sessions. The future model supports precise recommendations with reduced state space complexity. Prediction models are based on web log data that matches user actions. This prediction needs the detection of sequential web user access patterns and the use of these patterns to more accurately predict future user access.

THEORITICAL BACKGROUND

Web mining is defined as the application of data mining techniques aimed at discovering and extracting hidden information from data stored on the Internet. Another important goal of web crawling is to provide a mechanism to make data access more efficient and adequate. One interesting approach is to find out what information can be derived from user actions stored in log files. Web crawling can be divided into two types of approaches, namely the Information Retrieval Approach and the Database Approach. Information Retrieval Approach is used to aid and improve the search for information for users based on suggested or requested user profiles. On the other hand, Database Approach is used to model data on the web and integrate it [3].

Web crawling(mining) can be divided into three areas: Web Content Crawling, Web Structure Analysis and Web Usage Analysis. Web Content Crawling focuses on discovering/extracting useful information from Web Content/data/ documents, while Web Structure Analysis focuses on discovering how to mode lunderlying link structures of basic Internet links. Sometimes the distinction between the two is not very clear. Internet Usage Mining describes techniques that detect a user's usage pattern and attempt to predict user behavior [4].

WEB USAGE MINING

Web usage mining (Internet usage mining) is defined as an application that uses data mining to analyze and discover interesting patterns of user's usage data on the Internet. Usage data records the behavior of the user while browsing or carrying out transactions on a website. Typically,usage data arises from an Extended Common Log Format (ECLF) [3]. Internet usage information reported in the log is collected daily by all servers around the world. Microsoft Internet Information Server (IIS) creates a web log file [5].

Log files provide a list of page requests made to a given web server, where the request is characterized by, at least, the IP address of the computer that sent the request, the date and time of the request, and the URL of the page request. From this information, it is possible to reconstruct a user'sbrowsing sessions in a web site, where a session consists of a sequence of web pages viewed by the user in a given window of time. A web site

owner can use internet usage data mining techniques to gain insight into user behavior while visiting a website and use that knowledge to improve the design of the website [6].

MARKOV MODELS FOR FORECASTING USER'S ACTIONS

Techniques derived from Markov models have been widely used to predict the action that a user will take next. For this type of problem, Markov models are denoted by three parameters<A,S,T>, where A is a set of all possible actions that can be worked by the user; S is the set of all possible states for which the Markov model is constructed; and T is |S| × |A| Transition Probability Matrix (TPM), where each entry ti j corresponds to the probability of executing action j when the process is in state i. Typically, the input for these problems is the sequence of web pages that the user has accessed, and the goal is to create Markov models that can be applied to model and estimate the web page that the user is on more likely to access following. The main idea of the Markov model is to predict the next action based on the outcome of the previous actions. According to the web forecast, the next action is the forecast for the next page to visit. The previous actions correspond to the previous pages already visited. In web prediction, the K-order Markov model is the probability that a user visits the kth page given that she has visited the k-1 ordered pages. have already been visited[7]. For example, in first-order models, the state will be one page, while in second-order models, the state will only be two previously visited web pages. First order Markov models are used to model the sequence of pages wished by the user to predict the next page to be accessed. User actions are clustered by studying a mixture of first-order Markov models using association rules and ambiguous rules with Expectation Maximization levels. The activity of a random sample of users in each cluster is displayed, along with the size of each cluster. The example of web traffic is also applied on Myanmar-based travel agency website.

ASSOCIATION RULES

When shopping in a supermarket for a given set of transactions, where each transaction is a set of items, the association rule is X => Y, where X and Y are sets of items. The intuitive meaning of this rule is that database transactions that contain items in X tend to also include items in Y. For example, 98% of customers who buy auto accessories and accessories also buy car services; 98% is called the confidence of the rule. Support for X Rules => Y is the percentage of transactions that include both X and Y. Association rule generation can be applied to link pages that are most often referenced together in the same server session. Association rules can also be used as a heuristic for document prefetching to decrease a user's perceived latency when loading a page from a remote site. These rules are used to detect correlations between pages that are shared together during a server session. These rules show possible relationships between pages that are often seen together, even if they are not directly related, and can relate associations between groups of users with specific interests [8].

A transaction is a projection of a portion of the access log. Rules of the form Di → Dj are built, where Di and Dj are documents (URLs). An intuitiveinterpretation of these rules is that document C is likely to be requested by the same user sometimes after document Di has been requested, and there is no further request between requests for Di and Dj, because this usually happens according to journal, the media count is done differently because the order of the documents is taken into account. The trust is the support (Di Dj) / support (Di) ratio, the support (Di) is the total number of occurrences of the document Di in a transaction compared to the total number of transactions, and the support (Di Dj) is the total number of occurrences of the sequence Di Dj in transactions by total number of transactions. Only sequential subsequences of a custom transaction are supported. For example, the ABCD custom transaction supports the subsequences: AB, BC, and CD [8].

There are many techniques based on association rules for predicting web requests. The analysis is based on two-dimensional structures and weblogs are operated as data for training and testing. The first dimension is called the antecedent of the rules. The earlier rules make up the left side of the association rules. The second dimension is called the prediction rule criterion. Consequent prediction rules combine the confidence and support of the rule to form a single measure of choice. Prediction rules create the right side of association rules. Only the highest probability of consequent is taken into account. In this work, association rules are used to represent prediction rules.

AMBIGUOUS RULES

The main problem with association rules when applying large sets of data items is the detection of a huge number of rules and the trouble of framing which leads to a correct prediction. There are many different methods of creating association rules based on prediction models using blogs, but there are still some ambiguous rules. To solve this difficulty, the use of the ambiguity rule preserves both low state complexity and high precision of the outcomes.This prediction requires detecting sequential web user access
patterns and using these patterns to predict future user access. The first order Markov model, the second order Markov model, association rule analysis and ambiguous rules are used to build a predictive model.Entry to model training consists of web sessions, each includingthe sequence of pages that the user accesses during his visit to the site [9].

The sequential association rules of the form LHS → RHS are extracted from the session. Support and Trust are defined as follows [8]:

Supp = (count(LHS→RHS))/(number of sessions)

Conf = (count(LHS→RHS))/(count(LHS))

Consider the set of user web sessions in Table 1.YC, BC, MdC, NC, MgC, TC and ICareassigned to the names of web pages. Table 1 examines the following 5 user web sessions.

The supported values for web user sessions are YC = 2, BC = 5, MdC = 3, NC = 4, MgC = 4, TC = 3 and IC = 3.In the example of Latest Substring Association (LSA) rule from Table 2 is shown in Table1.

Table 1: Latest Substring Association rule

Antecedent (Previous)	Consequent (Next)	LSA
YC, BC, MdC, NC	MgC	NC ->MgC
BC, MdC, NC	MgC	
YC, BC, MdC, NC, MgC, TC	IC	TC -> IC
NC, MgC, BC, TC	IC	
BC, TC	IC	

Table 2: User web sessions from the web log file

WS1	YC, BC, MdC, NC, MgC
WS2	YC, BC, MdC, NC, MgC, TC, IC
WS3	BC, MdC, NC, MgC
WS4	NC, MgC, BC, TC, IC
WS5	BC, TC, IC

Table 3 shows that the support for the different collation sequences is calculated based on this training set. The row denotes the page go to previously and the column denotes the next page visited. Each field in the matrix is created by looking at how many times the web page is in a horizontal row and then the web page is in a vertical row.In this example, the web user sessions, web page MgC and web page NC are simultaneously from user sessions WS1 and WS3, and therefore web page (MgC, NC) supports 2 [9].

Table 3: Example of retrieved web access sequences and their number of supported 1st order Markov models

The following step is to create prediction rules from the remaining sequences. All the other sequences create prediction rules using the confidence level with the principle of maximum likelihood.The state probabilities of all sequences are computed and ranked. For example, given the history that the web page is BC, the probability of the condition is that BC->MdC is 60%. It is computed by dividing the sequence support value (BC, MdC) by the web page support value (BC); (3/5 = 60%). From the training, if the previous webpage is BC, the only impacts could be MdC and TC with their confidence levels of 60% and 40%. In this case, BC->MdC has the highest probability value. Then the prediction rule with the most probable state is selected.

Using the first-order Markov models to the user training sessions above, the most frequent Md state has occurred 3 times. Hence, the prediction rule BC->MdC is chosen because it is an unambiguous prediction. Using Markov models, the support value can decide that there is a 50/50 chance(ambiguous prediction) that the future page will be available to the user after accessing the page MgC can be BC or TC. Clearly, this information alone does not supportreal predictions on the future page that the user should go to.

Rule-selected		Confidence
Antecedent (previous)	Consequent (next)	
YC	BC	2/2=100%
BC	MdC	3/5=60%
MdC	NC	3/3=100%
NC	MgC	4/4=100%
MgC	BC	1/4=25%
TC	IC	3/3=100%
IC	-	0%
YC, BC	MdC	2/2=100%
BC, MdC	NC	3/3=100%

MdC, NC	MgC	3/3=100%
NC, MgC	BC	1/4=25%
MgC, BC	TC	1/1=100%
TC, IC	-	0%

The highest status probability was obtained for both pages using BC and TC. If the match rule (MgC->BC or MgC->TC) provides an ambiguous prediction, select the ambiguous rule with the highest RHS frequency (supported value) in the training weblogs to make the prediction.To pause the link and determine which page will lead to the most precise prediction, the ambiguous rule is selected on the right side (RHS), which is the most modern page in instruction blogs (webpage Ba = 5, webpage Ta = 3). Thus, the prediction rule Mg->Ba is selected. Considering the first order models, the second order models are also applied (Table 4). In this article, the support value is chosen when answering ambiguous predictions.It can be applied to all Markov models and association rules. This method prevents the complexity of the higher order Markov model. This method also improves the efficiency of Internet access predictions.

Table 4: Example of retrieved web access sequences and their number of supported 2nd order Markov models

2nd order	Support value count (frequency)						
Second item in sequence	YC	BC	MdC	NC	MgC	TC	IC
First item in sequence							
YC ->BC	0	0	2	0	0	0	0
BC ->MdC	0	0	0	3	0	0	0
MdC -> NC	0	0	0	0	3	0	0
NC ->MgC	0	1	0	0	0	1	0
MgC ->BC	0	0	0	0	0	1	0
TC ->IC	0	0	0	0	0	0	0

The result of the association rules generated and the ambiguous rules using Markov models as well as the corresponding confidence values are shown in Table 5.

Table5: Generation of rules using association rules and ambiguous rules for 1st and 2nd order Markov models.

Among all the rules generated, the rules that satisfy the highest confidence threshold (for example, 100%) are considered hard rules. Using these strict rules, the predictions of the web pages were made in such a way that the left side of the ruler was considered the visited web page and the right side of the rulers was the next web page.

Table 6: Result of strong rules generated (which contain highest confidence level of 100%) of all association rules above and ambiguous rules using Markov models

Rule-selected		Confidence probability
Antecedent (previous)	Consequent (next)	
YC	BC	2/2=100%
MdC	NC	3/3=100%
NC	MgC	4/4=100%
TC	IC	3/3=100%
YC, BC	MdC	2/2=100%
BC, MdC	NC	3/3=100%
MdC, NC	MgC	3/3=100%
MgC, BC	TC	1/1=100%

1st order	Support value count (frequency)						
Second item in sequence	YC	BC	MdC	NC	MgC	TC	IC
First item in sequence							
YC	0	2	0	0	0	0	0
BC	0	0	3	0	0	2	0
MdC	0	0	0	3	0	0	0
NC	0	0	0	0	4	0	0
MgC	0	1	0	0	0	1	0
TC	0	0	0	0	0	0	3
IC	0	0	0	0	0	0	0

This is an example of recommendation of popular link for Travels and Tours agency website.

pageYC www.travelsandtoursagency.com
page BC page YC
page NC page MdC
page MgC page NC
page IC page TC
page MdC page YC, BC
pageNC page BC, MdC
page MgC page MdC, NC
page TC page MgC, BC
page YC = Yangon City
page BC = Bagan City
page MdC = Mandalay City
page NC = Naypyitaw City
page MgC = Magway City
page TC = Taunggyi City
page IC = Innlay City

From the table above, you can easily see that the number of generated association rules and ambiguous rules using Markov models is too large, that the number of strong association rules and ambiguous rules using of Markov models is very small in both cases. Thus, the precision value for all association and ambiguity rules was calculated according to equation (1).

Accuracy = S/A

Where S = total number of strong association rules and ambiguous rules using Markov models, A = total number of association rules and ambiguous rules using Markov models. The precision value for all the association rules mentioned above and the ambiguous rules using Markov models have been evaluated in Table 7.

Table 7: Result of the precision value for all association rules and ambiguous rules using Markov models and the latest substring association rule

In the above reference, the LSA precision value is much better than the other two Markov models because it takes into account not only order and contiguity, but also relevance of web session information.

Rule-selected	A	S	Accuracy
1st order	6	2	(4/7) *100%=57%
2nd order	5	2	(4/6) *100%=67%
LSA	2	2	(2/2)*100%=100%

CONCLUSION

Markov models are a popular approach for predicting which web pages will be visited next. By using first and second order Markov models to predict user browsing sessions, the ambiguous rules of the problem can be resolved. In this paper, the value of the right side of the support for a survey can be determined by solving an ambiguous prediction in the first and second order Markov models. This technique can support improved prediction of network access than simply using separate legacy models separately. It can also be applied to all Markov models, association rules and more collective Markov models. This technique helps you better understand visitor surfing behavior and better respond to visitor requests (by reducing the length of your organization's navigation paths). This means that the system will predict where the most visitors will go. By using this method, the information can be used to help visitors find the information they need in an organization effectively and efficiently. For this reason, business owners may recommend that visitors visit popular places.

REFERENCES

1. https://www.wikipedia.org.com
2. Yang, Q., Li, T., Wang, K., "Building Association Rules Based Sequential Classifiers for Web Document Prediction", Journal of Data Mining and Knowledge Discovery, Netherland: Kluwer Academic Publisher, vol. 8, 2004, 253-273.
3. Yu Ya Win, "Prefetching Based-On Web User Clustering", PhD thesis, University of Computer Studies, Yangon 2007.
4. Yan Wang, "Web Mining and Knowledge Discovery of Usage Patterns", February 2000.
5. Rainer Gerhards, "Remotely Monitoring IIS Log Files".
6. Jose' Borges, "An Average Linear Time Algorithm for Web Usage Mining", University of London, September 2003.
7. Anand Charpate, Chetan Bramhankar, Prashant Gaikawad, A.D.Londhe, " Prediction of Link and Path for User's Web Browsing Using Markov Model" , International Journal of Computer Science and Mobile Computing , Pune University, India, .vol. 4, Issue. 2, February 2015, pg. 144-148.
8. Siriporn Chimphlee, Naomie Salim, Mohd Salihin Bin Ngadiman, Witcha Chimphlee, "Using Markov Model and Association Rules for Web Access Prediction", Advances in Systems, Computing Sciences and Software Engineering, 2005.
9. Thanakorn Pamutha, Chom Kimpan, Siriporn Chimphlee, Parinya Sanguansat, "Improving Web Page Prediction Using Default Rule Selection", International Journal of Advanced Computer Science and Applications, Vol. 3, No.11, 2012.