Implementing Prediction System by Using Ontology based PHS Algorithm

Nu War Hsan, Sabai Phyu
University of Computer Studies, Yangon
nuwarhsan@ucsy.edu.mm, sabaiphyu72@gmail.com

Abstract

Web prediction is a classification approach to predict the next step of Web pages that a user may visit on the knowledge of the previously visited pages. The web usage mining techniques are used to analyze the web usage patterns for a web site. The user access log is used to fetch the user access patterns and these are used in the prediction process. There are many methods based on web usage mining for prediction system but most of all are based on traditional methods such as sequential pattern mining, clustering and so on. In this paper, the ontology based PHS (Perfect Hashing and Shrinking) algorithm is used in developing list by applying semantic similarity related with the domain ontology.

Keywords: *Semantic Web, Ontology based PHS, Semantic Similarity, Web Server Log Files, Domain Ontology, Reference Ontology, PHS (Perfect Hashing and Shrinking), PHP (Perfect Hashing and Pruning), DHP(Direct Hashing and Pruning).*

1. Introduction

Finding relevant information on the Web has become a real challenge. This is partly due to the volume of data available and to the lack of structure in many Web sites. In recent years, large amount to informative websites, web pages and web documents are popular as huge collection. Any popular search engine returns thousands of related links to a search query. But it has become difficult for users to get the most relevant information from the related information efficiently. Modeling and analyzing web navigation behavior is helpful in understanding what information user's online demands.

Without such semantic knowledge, personalization system cannot predict different types of complex objects based on their underlying properties and attributes. The integration of semantic knowledge is the primary challenge for the next generation of personalization systems. In this paper, the integration of semantic information is drawn from a web's application domain knowledge into all phases of web usage mining process. The goal of proposed system is to have an intelligent semantics-aware web mining framework. In this paper, ontology based PHS (Perfect Hashing and Database Shrinking) is used to generate frequent item sets and semantic association rules. And then the correlation based similarity is used to compare the user current navigation paths and offline navigation paths to produce the prediction.

2. Related Work

Om Kumar and P.Bhargavi [1] discussed about the log files and uses Web mining techniques to extract usage patterns by using WEKA. In this paper, classification into three domains of web mining is explained. And then the types of server logs are widely described. And all the log formats specified above has fields that record http request in the form of elapse time and response in the form of status code are explained in detail. And it also described as the extended work to mine the log file based user clicks.

Akshay Kansara and Swati Patel [3] presented the combination of the classification and clustering techniques to predict user future movements. In this proposed system, a clustering algorithm is used to discover the navigation patterns. The experimental results prove that the proposed amalgamation of techniques is efficient both in terms of clustering and classification.

Dilpreet Kaur and Sukhpreet Kaur [2] presented an overview of past and current evaluation in user future request prediction using web usage mining. The approach is based on the new graph partitioning algorithm to model user navigation patterns for the navigation patterns mining phase. LCS algorithm is used for classifying current user activities to predict user next movement. It also described as the future work to predict user's future requests by using different techniques such as classification, clustering and association rule mining.

Mehrdad Jalali, Norwati Mustapha, Ali Mamat and Md.Nasir B Sulaiman [4] described the testing of their proposed model to improve the quality of prediction results. LCS algorithm is used in this system to achieve more accurate recommendation for long patterns of the current user activities in the particular web sites. And they also described future work to take into account the semantic knowledge about underlying domain to improve the quality of the recommendation. And it also described the integrating semantic web and web usage mining can achieve best recommendation in the dynamic huge web sites.

3. Theory Background

3.1 Web Usage Mining

Web usage mining focuses on techniques that could predict user behavior while the user interacts with the Web. The web usage mining can be classified into three processes, consisting of the data preparation, pattern discovery and pattern analysis phases. In the first phase, web log data are preprocessed in order to identify users, sessions, page views, and so on. In the second phase, statistical methods, as well as data mining method (such as association rules, sequential patterns discovery, clustering and classification) are applied in order to detect interesting patterns. These patterns are stored so that they can be further analyzed in the third phase of the Web usage mining process.

4. Overview of Proposed System

There are four steps in this proposed system. The first step is preprocessing the logs (raw data) based on user identification and session identification according to the user session and visit. The second step is building the reference ontology according to the web site structure and updates the domain ontology by using domain information of user transaction file. The third step is applying the ontology based PHS algorithm to develop the semantic association rules. Finally, the calculation of correlation based similarity is applied to predict the next pages for current user session.

4.1 Web Logs Pre-processing Methods

The process of extracting web usage logs to implement prediction system performs three main steps: Data preprocessing, using ontology based PHS algorithm and Prediction. The first step and the second are performed off-line and the last is online. The web server registers the entire request made to the server in the log file including request time, request type (GET, POST and HEAD), http version, user agent information, client IP address, response status and referrer address.

Firstly, the non-responded requests are pruned from the status field of the log entry. The second step is to eliminate the requests made by software agents which sometimes automatically request web content from a web site. The third step is to remove the irrelevant requests from the log file such as image request or style sheet request which are not taken into account since these files are auxiliary files for displaying web site to the user. To sum up, the status code is used to prune the non responded requests, address fields are used to prune Web page addresses and user agent fields are used to prune the web crawlers. The next step after pruning is the extraction navigation history of each session from log files. The navigation history is the set of web objects requested by the user in his active session time. A session is established when the end-user makes the first request to the web server and the session is torn down after a period of idle time from end-user.

4.2 Building Reference Ontology and Domain Ontology according to user transaction

Firstly, the reference ontology is prebuilt related with education domain. When the user transaction is produced, the domain ontology is built according to link structure of web site from these transactions.

4.3 Ontology based PHS Algorithm

In this proposed system, the PHS algorithm is used with some modification to map with ontology instances to generate the semantic association rules. The two algorithms namely Semantics-Aware Framework and Semantic-Aware PHS are applied to generate the semantic frequent semantic objects and semantic association rules using semantic similarity between concepts.

In Semantic-Aware Framework, the input patterns are clean web logs, domain ontology, maximum semantic distance which is specified by the user according to the relationship of the domain side and minimum support count. And then the clean web logs are mapped with semantic objects and build the semantic matrix to calculate semantic distance of each other.

After generating the semantic matrix, the Semantic Aware PHS is called to generate the frequent objects and semantic association rules. The Semantic Aware PHS algorithm is used the framework of PHS algorithm combined with the calculation of semantic distance to produce the accurate results for prediction. At the beginning of the Semantic Aware PHS algorithm, scanning and counting the support, filtering out the items with right support and dropping the remaining which does not have enough support from the database. After that, hash table for candidate two item sets is constructed where semantic distance of each pair is less than the specified semantic distance. When this work has been done, the table contained candidate 2-itemsets and the number of their occurrence are generated.

From next step to final, the algorithm is different to other sequential pattern mining algorithms. The frequent two item sets are unique corresponding to a word in a new system. This may also called encoding such that item set AB assign as a word namely "a" and another CD assigns as "b". And the items which are not frequent two item sets are removed from database. The three item sets will be in form of two item sets. For example: joining with a (AB) and b (BC) will result in (ab).

By this process, the database reduces, as all the non frequent item sets will be removed during processing. From theoretical point of view, it is better on the side of execution time because it does not only eliminate collision but also fix the length of candidate items sets to simplify the task of making hash function.

To calculate semantic similarity between two concepts, the following concept similarity method is applied based on semantic distance of each concept.

4.3.1. Concept Similarity Matching Method Based on Semantic Distance

The algorithm takes two concepts as input and computes a semantic similarity as output in four steps:

Step 1: **Weight Allocation**: Determine weight according to the relationship between root node of ontology and other concept nodes.

Step 2: **Node routing table generation**. Record all paths between node and concept nodes, generate routing table.

Step 3: Semantic distance computation. Compute semantic distance according to the node routing table. The semantic distance is the sum of weight between the concepts that have the inheritance relationship.

Step 4: Semantic similarity computation. Construct similarity function, and compute semantic similarity between concepts based on semantic distance.

4.4 Correlation based Similarity and Prediction

In this proposed system, the correlation based similarity algorithm is used to map with the previous navigation paths and user current navigation paths to generate the prediction

$$sim(i,j)=\frac{\sum_{u\in U}(R_{u,i}-R_i)(R_{u,j}-R_j)}{\sqrt{\sum_{u\in U}(R_{u,i}-R_i)^2}\sqrt{\sum_{u\in U}(R_{u,j}-R_j)^2}} \quad (1)$$

where sim(i,j) = similarity of current user navigation paths and offline navigation paths;

$R_{u,i}$ = rating of user's current navigation paths

R_i = average rating of the current navigation paths

$R_{u,j}$= rating of offline navigation paths

R_j = average rating of the offline navigation paths

$$R_{i,j}=\frac{P(t\mid j)}{\sum_{j=1,\dots n}P(t\mid j)} \quad (2)$$

where $R_{i,,j}$= rating of each transaction of user navigation paths. (1)

$$R_{i,j}=\frac{P(\mathrm{P(t\ in\ j)})}{\sum_{j=1,\dots n}freq\,(\mathrm{P(t\ in\ j)})} \quad (3)$$

freq (t in j) = occurrence of transaction t in user navigation paths.

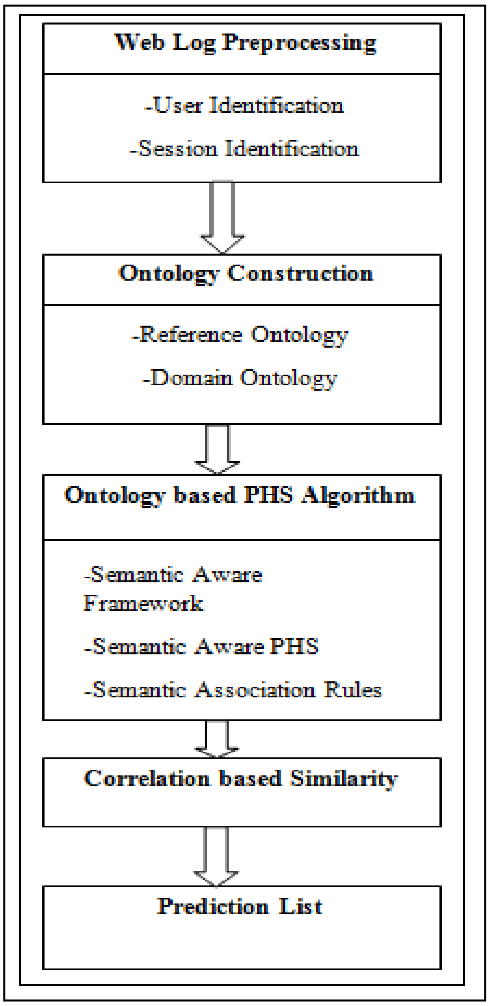
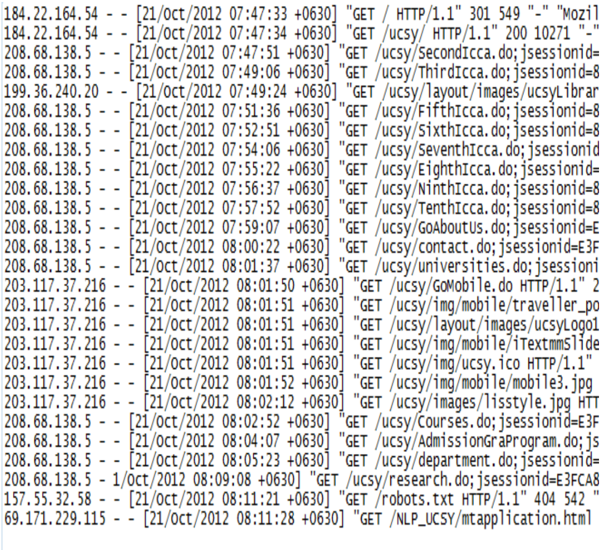


Figure 1: Steps of Proposed System

5. Illustration and System Design of the Proposed System

In this proposed system, the log files from the web site (http://www.ucsy.edu.mm) are applied.

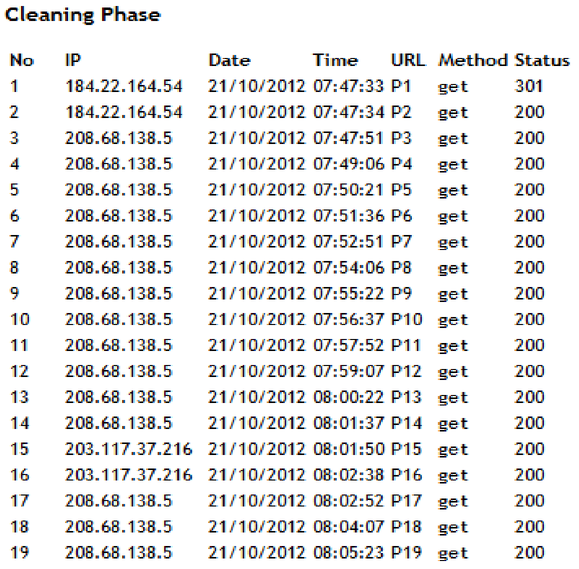
And then the preprocessing methods are applied to raw log files and clean log files are generated after preprocessing. For getting the user name, the user login information (Date, User ID, Client IP, In-time, Out-time) attributes can be user the user identification step.



```
184.22.164.54 - - [21/Oct/2012 07:47:33 +0630] "GET / HTTP/1.1" 301 549 "-" "Mozil
184.22.164.54 - - [21/Oct/2012 07:47:34 +0630] "GET /ucsy/ HTTP/1.1" 200 10271 "-"
208.68.138.5 - - [21/Oct/2012 07:47:51 +0630] "GET /ucsy/SecondIcca.do;jsessionid=
208.68.138.5 - - [21/Oct/2012 07:49:06 +0630] "GET /ucsy/ThirdIcca.do;jsessionid=8
199.36.240.20 - - [21/Oct/2012 07:49:24 +0630] "GET /ucsy/layout/images/ucsyLibrar
208.68.138.5 - - [21/Oct/2012 07:51:36 +0630] "GET /ucsy/FifthIcca.do;jsessionid=8
208.68.138.5 - - [21/Oct/2012 07:52:51 +0630] "GET /ucsy/SixthIcca.do;jsessionid=8
208.68.138.5 - - [21/Oct/2012 07:54:06 +0630] "GET /ucsy/SeventhIcca.do;jsessionid
208.68.138.5 - - [21/Oct/2012 07:55:22 +0630] "GET /ucsy/EighthIcca.do;jsessionid=
208.68.138.5 - - [21/Oct/2012 07:56:37 +0630] "GET /ucsy/NinthIcca.do;jsessionid=8
208.68.138.5 - - [21/Oct/2012 07:57:52 +0630] "GET /ucsy/TenthIcca.do;jsessionid=8
208.68.138.5 - - [21/Oct/2012 07:59:07 +0630] "GET /ucsy/GoAboutUs.do;jsessionid=E
208.68.138.5 - - [21/Oct/2012 08:00:22 +0630] "GET /ucsy/contact.do;jsessionid=E3F
208.68.138.5 - - [21/Oct/2012 08:01:37 +0630] "GET /ucsy/universities.do;jsessioni
203.117.37.216 - - [21/Oct/2012 08:01:50 +0630] "GET /ucsy/GoMobile.do HTTP/1.1" 2
203.117.37.216 - - [21/Oct/2012 08:01:51 +0630] "GET /ucsy/img/mobile/traveller_po
203.117.37.216 - - [21/Oct/2012 08:01:51 +0630] "GET /ucsy/layout/images/ucsyLogo1
203.117.37.216 - - [21/Oct/2012 08:01:51 +0630] "GET /ucsy/img/mobile/iTextmmSlide
203.117.37.216 - - [21/Oct/2012 08:01:51 +0630] "GET /ucsy/img/ucsy.ico HTTP/1.1"
203.117.37.216 - - [21/Oct/2012 08:01:52 +0630] "GET /ucsy/img/mobile/mobile3.jpg
203.117.37.216 - - [21/Oct/2012 08:02:12 +0630] "GET /ucsy/images/lisstyle.jpg HTT
208.68.138.5 - - [21/Oct/2012 08:02:52 +0630] "GET /ucsy/Courses.do;jsessionid=E3F
208.68.138.5 - - [21/Oct/2012 08:04:07 +0630] "GET /ucsy/AdmissionGraProgram.do;js
208.68.138.5 - - [21/Oct/2012 08:05:23 +0630] "GET /ucsy/department.do;jsessionid=
208.68.138.5 - 1/Oct/2012 08:09:08 +0630] "GET /ucsy/research.do;jsessionid=E3FCA8
157.55.32.58 - - [21/Oct/2012 08:11:21 +0630] "GET /robots.txt HTTP/1.1" 404 542 "
69.171.229.115 - - [21/Oct/2012 08:11:28 +0630] "GET /NLP_UCSY/mtapplication.html
```

Figure 2. Raw Log Files in Text Format

To group the activities of a single user from web log files is called a session. As long as user is connected to the website, it is called the session of that particular user. In this proposed system, 30 minutes is assigned as a session time-out. In Figure 1, raw log files of proposed system are described in text format.



Cleaning Phase

No	IP	Date	Time	URL	Method	Status
1	184.22.164.54	21/10/2012	07:47:33	P1	get	301
2	184.22.164.54	21/10/2012	07:47:34	P2	get	200
3	208.68.138.5	21/10/2012	07:47:51	P3	get	200
4	208.68.138.5	21/10/2012	07:49:06	P4	get	200
5	208.68.138.5	21/10/2012	07:50:21	P5	get	200
6	208.68.138.5	21/10/2012	07:51:36	P6	get	200
7	208.68.138.5	21/10/2012	07:52:51	P7	get	200
8	208.68.138.5	21/10/2012	07:54:06	P8	get	200
9	208.68.138.5	21/10/2012	07:55:22	P9	get	200
10	208.68.138.5	21/10/2012	07:56:37	P10	get	200
11	208.68.138.5	21/10/2012	07:57:52	P11	get	200
12	208.68.138.5	21/10/2012	07:59:07	P12	get	200
13	208.68.138.5	21/10/2012	08:00:22	P13	get	200
14	208.68.138.5	21/10/2012	08:01:37	P14	get	200
15	203.117.37.216	21/10/2012	08:01:50	P15	get	200
16	203.117.37.216	21/10/2012	08:02:38	P16	get	200
17	208.68.138.5	21/10/2012	08:02:52	P17	get	200
18	208.68.138.5	21/10/2012	08:04:07	P18	get	200
19	208.68.138.5	21/10/2012	08:05:23	P19	get	200

Figure 3. Cleaning Phase of Raw Logs

And then the preprocessing step such as user identification and session identification are applied this raw log files and clean log files are produced which is described in Figure 2.

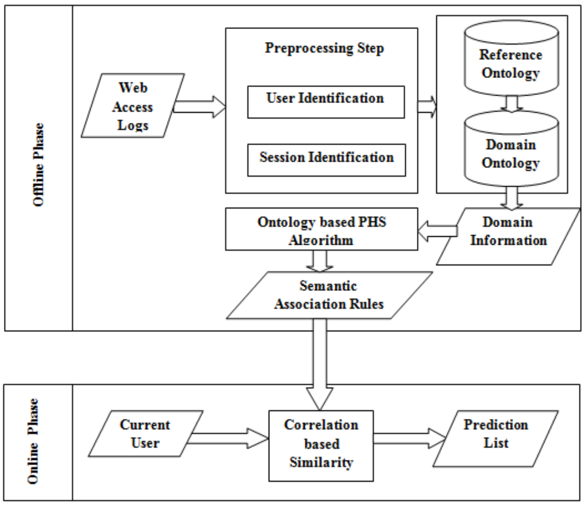


Figure 4. System Framework

The domain ontology is constructed with web site's structure and the keywords as concepts (related with education domain) and relationships are identified. When the clean web log files are mapped with ontology files and produce the semantic web logs files. These logs are mapped with domain ontology and calculate web page similarity using semantic distance. It is calculated by applying with the reference ontology which is prebuilt according to the education domain.

Ontology based PHS Algorithm is used according to the semantic distance and generates the semantic association rules. And then the correlation similarity is applied to generate the prediction list for with the current user's session and semantic association rules.

6. Experimental Results

In this proposed system, it is easy to understand advantages of PHS algorithm after studying the traditional methods such as Apriori algorithm, Sequential Pattern Mining Algorithm.

In addition, retrieving relevant items and transactions during the process can lead to the database to be reduced. Figure 4 shows that the difference of the execution time by using Apriori and PHS algorithm. The PHS algorithm which is better performance than Apriori because it is only reduced the number of scanning the whole database but also decrease the size of database.

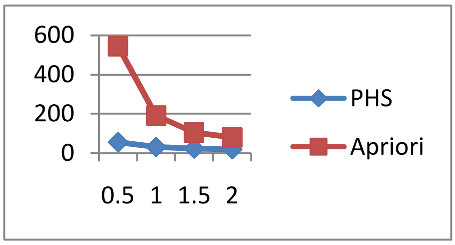


Figure 5. Comparison of Execution time of Apriori and PHS

7. Conclusion

In this paper, ontology based PHS algorithm is performed using domain ontology and reference ontology and generate the semantic association rules by using semantic similarity. It can be solved the weakness of previous algorithms such as DHP and PHP algorithm by eliminating collision and fixing the length of candidate item sets. This proposed system overcomes the drawbacks of traditional web usage mining such that the results are in the form of web pages without semantic meaning of common navigation profiles. By introducing the semantic information, web usage mining algorithms are performed in terms of ontology individuals instead of web pages.

This paper has provided analysis of web usage mining for browsing behavior of a user and subsequently to predict desired page research available. Web usage mining is fast rising area technology today generated log information can be useful in various ways. As future direction, we plan to develop other semantic web usage mining system by using other clustering algorithms.

References

[1] Om Kumar,,P.BHARGAVI June 2013 Analysis of Web Server Log By Web Usage Mining For Extracting Users Patterns, International Journal of Computer Science Engineering.

[2] Dilpreet kaur, Sukhpreet Kaur April 2013 A Study on User Future Request Prediction Methods Using Web Usage Mining, International Journal of Computational Engineering Research.

[3] Akshay Kansara, Swati Patel, May 2013 Improved Apporach to Predict user Future Sessions using Classification and Clustering, International Journal of Science and Research (IJSR).

[4] Mehrdad Jalai, Norwati Mustapha,Ali Mamt, Md Nasir B Sulaiman October,2009. A Recommender System for Online Personalization in the WUM Applications.
Antony Scime, Web Mining Applications and Techniques, State University of New York College at Brockport, USA.

[5] Hassan Najadat, Amani Shatnawi and Ghadeer Obiedat, Jordan University of Science and Technology, A New Perfect Hashing and Pruning Algorithm for Mining Association Rule, IBIMA Publishing.

[6] Kalyan Beemanapalli, October 2006, A Framwork for Incorporating Domain Information into Usage Mining Based Recommendations.

[7] M.Andrea Rodriguez and Max J.Egenhofer, Determining Semantic Similarity Among Entity Classes from Different Ontologies, IEEE Transactions on Knowledge and Data Engineering.

[8] Thabet Slimani, Description and Evaluation of Semantic Similarity Measures Approaches.

[9] Sampath P and Ramya D, Performance Analysis of Web Page Prediction With Markov Model, Association Rule Mining(Arm) and Association Rule Mining with Statistical Features(Arm-Sf), IOSR Journal of Computer Engineering.