DEVELOPING PREDICTION SYSTEM BY USING ONTOLOGY BASED PHS ALGORITHM

NU WAR HSAN

Ph.D Candidate, University of Computer Studies, Yangon
E-mail: nuwarhsan87@gmail.com

Abstract- As the World Wide Web is rapidly grown, the information sources are changed daily. Increasing amount of information is the vital problem for information retrieval. Web prediction is a classification problem in which we have to predict the next set of Web pages that a user may visit based on the knowledge of the previously visited pages. Predicting user's behavior can be developed in several applications in the internet environment. The web usage mining techniques in research area are used to analyze the web usage patterns for a website. 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 reference ontology is build according to the web site structure and the domain ontology is built according to the clean web log files. And then the domain information is generated and ontology based PHS algorithm is used in developing sequential pattern. And then, the prediction list is produced by applying semantic similarity related with the domain ontology.

Keywords- Domain Ontology, Perfect Hash Pruning, Semantic Similarity, Web Usage Mining.

I. INTRODUCTION

Although a lot of research has been done on Web Usage Mining, there are many researches that need to be extended. When the user browses the web pages, user leaves some valuable information in web log. This web log information is used as the input to find out navigation behavior of user. By applying this, the user's required information from which websites can be retrieved. Web mining is the application of data mining techniques to automatically retrieve, extract and evaluate information for knowledge discovery from web documents and services. The objective of Web Mining is vast, heterogeneous and distributing documents. Web mining is divided into three types. They are Web content mining, Web structure mining and Web usage mining. Web content mining deals with the discovery of useful information from the web contents or data or documents or services. Web content may be unstructured (plain text), semi-structured (HTML documents), or structured (extracted from databases into dynamic web pages).

Web structure mining mines the structure of hyperlinks within the web itself. Structure represents the graph of the link in a site or between the sites. The goal of web structure mining is to categorize the Web pages and generate information such as the similarity and relationship between them, taking advantage of their hyperlink topology.

Web usage mining mines the log data stored in the web server. Web usage mining aims to discover useful knowledge from hyperlinks, which represent the structure of the Web. Hyperlink is a link that exists in a web page or another web page. Finally, web usage mining, also acknowledged as Web Log Mining, aims to capture and model behavioral patterns and profiles of users who interact with a web site.

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. The main objectives of this proposed system is to use an existing built ontology intended to enhance information retrieval. And then to identify web user's profiles through an analysis of their visits with web usage mining method by using PHS (Perfect Hashing and Shrinking Database) algorithm.

II. RELATED WORK

Akshay Kansara and Swati Patel presented the combination of the classification and clustering techniques to predict user future movements. In their 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 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 and Norwati Mustapha discussed the interplay of the Semantic Web with Web usage mining and also gave a relation of two research areas. And in this paper, a framework of integrating semantic Web and Web usage mining is described. And it also

describes the future direction to develop a semantic Web usage mining system.

V Sitha Ramulu, Ch.N.Santhosh Kumar, K.Sudheer Reddy presented the combination of semantic web and web mining. And the steps of integrating domain ontology into web mining to become semantic web mining are described in detail. The semantic web mining concepts are also introduced.

Jike Ge and Yuhi Qiu provided as an experimental comparison of their methods against traditional similarity measures and proved the benefits of their approach. They introduced weight allocation, node routing table generation, semantic distance computation and semantic similarity computation are described in detail.

M.ahendra Thakur, Yogendra Kumar Jain and Geetika Silkari presented the Semantic Web Personalization, a personalized system that integrates usage data with content semantics, expressed in ontological terms, in order to compute semantically enhanced navigational patterns and effectively generate useful recommendations. And they also presented a novel approach for enhancing the quality of recommendations based on the underlying structure of a web site.

III. THEORY BACKGROUND

A. Semantic Web and Web Usage Mining

The semantic web is an extension of the current web in which information is given well-defined meaning, enabling computers and people to work in better cooperation. Semantic Web Usage Mining (SWUM) aims to integrate two research areas Semantic Web and Web Usage Mining to obtain more meaningful user behaviors in the Web environment. To better understand user's next intentions form observing him whole navigating a Website, all semantically interaction data needs to be tracked as well as tracking usage data. Semantic Web and Web Usage Mining can be integrated to have a rich semantic model of content and structure of a site. Deeper interaction of the Website's user with the site can be utilized semantic knowledge and the analyzing about usage patterns discovered in the mining phase can be done by using it. Moreover, it can improve the quality of predictions in the usage based system.

B. Web Usage Mining for Ontology Management

This section presents the presented methodology for supporting the management of Web site's ontology based on usage mining.

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.

Step 1: Web Logs Pre-processing Methods

In order to preprocess the logs for usage analyses, the recently methodologies for logs data preprocessing proposed by (Tanasa 2005).

- A URI is a compact string of characters for identifying an abstract or physical resource.
- A Web resource is resource accessible through any version of the HTTP protocol.
- A Web server is a server that provides accesses to Web sources.
- A Web Page is a set of data constituting one or several Web resources that can be identified by a URI.
- A page view occurs at a specific moment in time, when a Web browser displays a Web page.
- A Web browser or Web client is client software that can send Web requests, handle the responses and display requested URIs.
- A user is a person using a Web browser.
- A Web request is a request a Web client makes for a Web resource.
- A Web server log file contains Web requests made to the Web server, recorded in chronological order.
- A user session is a delimited number of a user's explicit Web requests across one or more Web servers.
- A visit is a subset of consecutive page views form a user session occurring closely enough (by means of a time threshold or a semantic distance between pages).

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.

Step 2: Building Reference Ontology and Domain Ontology according to user transaction
Firstly, the reference ontology is according to the web site structure. When the user transaction is produced, the domain ontology is built according to the domain information from these transactions. The domain ontology is built according to these following steps: extracting some keywords representative of the domain, finding a collection of websites related to the previous keywords, searching the initial keywords in a website and finding the preceding and following words (to these words are candidates to be concepts).

Step3: 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. 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 2 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 2 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 2-item sets are removed from database. The 3-itemsets will be in form of 2-itemsets. 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.

Step 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.

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

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

$$R_{i,j} = \frac{P\,(t \mid j)}{\sum_{j=1,\ldots\ n} P\,(t \mid j)} \qquad (1)$$

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

$$R_{i,j} = \frac{P(\mathrm{P(t\ in\ j)})}{\sum_{j=1,..n} freq\ (\mathrm{P(t\ in\ j)})} \qquad (2)$$

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

IV. ILLUSTRATION AND SYSTEM DESIGN OF THE PROPOSED SYSTEM

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

```
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
```

Fig 1: Raw Log Files in Text Format

And then the preprocessing methods are applied to raw log files and clean log files are generated after preprocessing. In this proposed system, each combination of IP address/Agent/Operating System is assumed as a single user. 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. 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 Fig 1, raw logs files of proposed system are described in text format. 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 Fig 2.

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

Fig 2: Cleaning Phase of Raw Logs

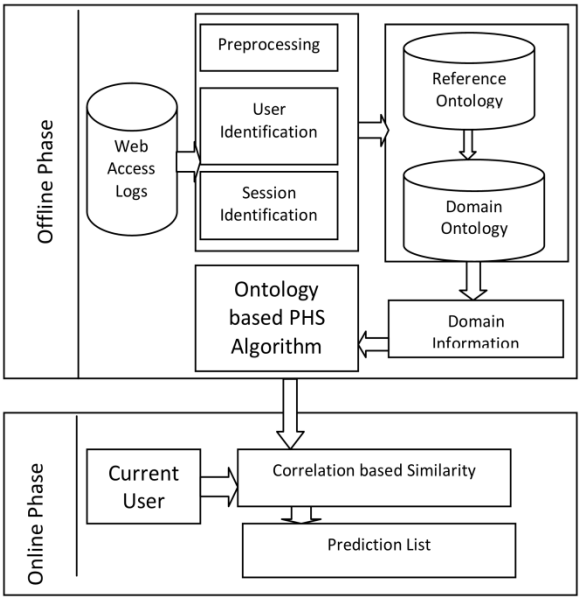


Fig 3: System Framework

The domain ontology is constructed with web site's structure and the keywords, concepts 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 by applying ontology based PHS algorithm and generate the semantic association rules using semantic similarity. And then the correlation based similarity is applied to generate the prediction list for with the current user's session and semantic association rules.

V. PERFORMANCE MEASURES

In this proposed system, the two methods are applied to measure the system's performance: Recall and Precision.

(A) Recall

The measure of how much relevant information the system has extracted (coverage of system).

$$\text{Re}\,call = \frac{\#\ of\ relevant\ links\ given\ by\ the\ system}{Toal\ \#\ of\ relevant\ links} \quad (3)$$

(B) Precision

Measure of how much of the information the system returns is correct (accuracy).

$$Precision = \frac{\#of\ relevant links given by the system}{Toal \#of\ links retrieved} \quad (4)$$

VI. 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. And also other hashing algorithm namely DHP (Direct Hash and Pruning) algorithm can reduce significantly the number of total scanning database by using hash mechanism.

In addition, trimming relevant items and transactions during the process can lead to the database to be smaller than smaller. Figure 4 shows that the difference of the execution time by using Apriori and DHP algorithm. The DHP 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. The disadvantage of DHP algorithm has collision in more than one item sets. The PHP (Perfect Hashing and Database Pruning) was solved this problem by designing two distinct item sets into two different places in hash table. But it has still disadvantages such as big size in hash table and it may face the problem of needing more parameters to control the function and it may be more complicated. And then PHS (Perfect Hashing and Data Shrinking) is introduced to solve the above issues and reduce the complex level of hash function.

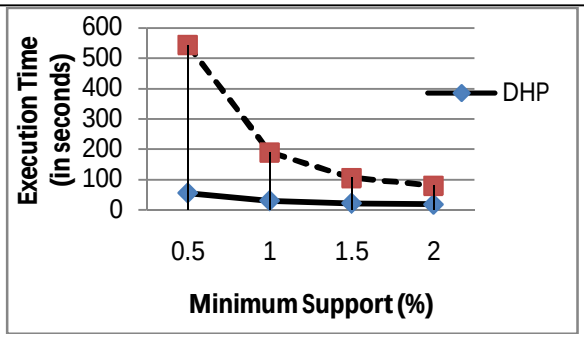


Fig 4: Comparison of Execution time of Apriori and DHP

An analysis study of Apriori, DHP (Direct Hashing and Pruning), PHP (Perfect Hashing and DB Pruning) and PHS(Perfect Hashing and Data Shrinking) in the execution of prediction systems are presented in Fig 5. The proposed system is tested with different set of log transactions. The analysis shows that the PHS algorithm can improve the prediction accuracy 27% more than others.

Table 1. Prediction Accuracy Analysis of Apriori, DHP, PHP and PHS Algorithms

Transaction	Apriori	DHP	PHP	PHS
100	0.423	0.525	0.635	0.699
200	0.436	0.543	0.647	0.703
300	0.465	0.564	0.653	0.724
400	0.478	0.578	0.665	0.735
500	0.489	0.593	0.673	0.747

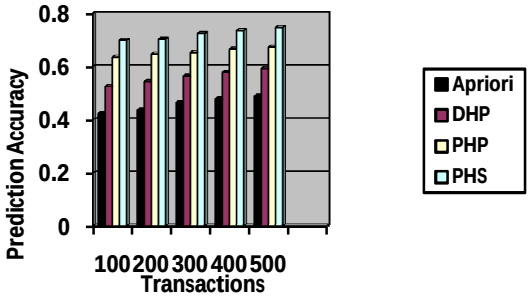


Fig 5: Prediction Accuracy Analysis- Apriori, DHP, PHP and PHS Algorithm

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] Antony Scime, Web Mining Applications and Techniques, State University of New York College at Brockport, USA.

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

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

[4] Jike Ge and Yuhui Qiu, 2008, Concept Similarity Matching Based on Semantic Distance, Fourth International Conference on Semantics, Knowledge and Grid.

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

[6] Mehrdad Jalali and Norwati Mustapha, Semantic Clickstream Mining, University Putra Malaysia.

[7] Mahendra Thakur, Yogendra Kumar Jain and Geetika Silakari, February 2011, Query based Personanlization in Semantic Web Mining, International Journal of Advanced Computer Science and Application.

[8] Suresh Shirgave and Prakash Kulkarni, October 2013, Semantically Enriched Web Usage Mining for Predicting User Future Movements, International Journal of Web and Semantic Technology.

[9] Tanasa, D. (2005). Web Usage Mining: Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support. PhD thesis, University of Nice Sophia Antipolis.

[10] V.Sitha Ramulu, Ch.N.Santhosh Kumar and K.Sudheer Reddy, July 2012, A Study of Semantic Web Mining: Integrating Domain Knowledge into Web Mining.

★★★