Semantic Web Usage Mining to Develop Prediction System

Nu War Hsan
University of Computer Studies Yangon
Myanmar

ABSTRACT
Technology innovation has led to an explosive growth of recorded information, with the Web being a huge repository under no editorial control. Providing with people with access to information is not the problem; the problem is that people with varying needs and preferences navigate through large Web structures, missing the goal of their inquiry. Web usage mining has been applied effectively in prediction system to overcome deficiencies of traditional approaches. The traditional approach does not take into account the semantic knowledge about the underlying domain. Prediction system cannot predict different types of objects based on their underlying attributes and properties. The integration of primary knowledge is the primary challenge of prediction system. In this paper, the prediction system is developed which extracts the key web objects from web log file and apply a semantic web to mine actionable intelligence.

General Terms
Web Usage Mining, Semantic Web, Ontology, Hashing Algorithm.

Keywords
Semantic Web, Ontology based PHS, Semantic Similarity, Web Server Log Files, Domain Ontology, Reference Ontology, PHS, PHP, DHP.

1. INTRODUCTION
In recent years, large amount of 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. It can be used for different purposes such as personalization, recommendation system improvement and site. Web usage mining is concerned with finding user navigational patterns on the World Wide Web by extracting knowledge from web logs.

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
Mehrdad Jalali and Norwati Mustapha [5] 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.

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.

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. Brigitte Trousse, Marie_Aude, Benedicte Le Grand.Yves Lechevallier and Florent Masseglia proposed the usage analysis of the chosen Web site in complement of the existing approaches based on content analysis of Web pages. It described as the main fact that web mining can be useful to add semantic annotations (ontology) to Web documents and to populate these ontological structures.

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.

3.2 Semantic Web and Web Usage Mining

The semantic web is an extension of the current Web in which information in 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 from observing him while 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.

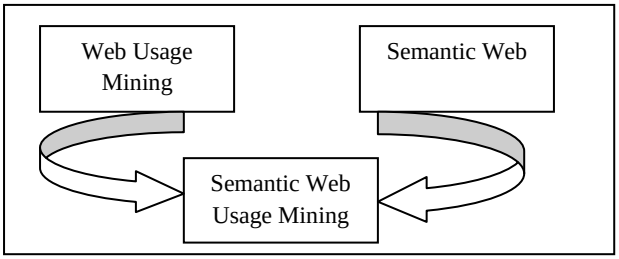

Fig 1: Semantic Web Usage Mining

3.3 Framework for Prediction System

The aim of prediction based on web usage mining is to predict a set of objects to the current user as determined by matching patterns. This task is performed by matching the current user's actions with the usage patterns discovered through web usage mining. The prediction phase is performed as the online component of this proposed system. 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 on-line.

Normally, 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.

The next step of pre-processing is to map between ontology individuals and the requested Web address in the Web server log. The Web server does not registers semantic information about the request in the log file, only the address of the request. Therefore, before starting the frequent sequence finding, mapping between ontology and the Web site address is carried out. To include semantic information on a Web page, there will be an ontology which defines the classes of the domain space and their properties showing the relationships among them.

Some of the traditional prediction systems have the problems of scattered and unstructured information that does not match users' expectations. This proposed system develops the solution to this problem by the following. Firstly, the system domain is developed by using reference ontology intended to enhance information retrieval. Secondly, the web users' profiles are identified through an analysis of their visits with ontology based PHS algorithm. The domain ontology is updated with extracted knowledge. The prediction result is produced for the user to adapt with each individual preference according to his profile.

3.3.1 Definition of Components in Semantic Aware Framework

A core ontology is defined as a structure O: $<C, \leq_C, R, \leq_R)$ consisting of: two disjoint sets C and R whose elements are called concept identifiers and relation identifiers, respectively. A partial order $\leq_C$ on C is called concept hierarchy or taxonomy and a partial order $\leq_R$ on R is called relation hierarchy. A semantic object o_i is represented as a tuple $<pg, ins_i>$. The term **pg** represents the web page which contains the object/product, usually an URL address of the page. The term $\mathbf{ins_i}$ is an instance of a class $c \in C$ from the reference ontology O that represents the product being referenced, where i is an index of an enumeration of the objects in the sequence, from the web access sequence database being mined. During preprocessing, a simple parser goes through the web log, extracts all ontology instances represented by web pages in the log and converting the web log to a sequence of semantic objects.

The semantic distance Matrix M is an n*n matrix of all the semantic distance between the n objects represented by web pages in the sequence database. **Maximum semantic distance d** is a value which represents the maximum allowed semantic distance between any two semantic objects. Maximum semantic distance can be user-specified or can be calculated in equation: d=min_sup*|R|.

3.3.2 Semantic Aware Framework

Input: clean web log WL={w_1,w_2,…w_n};Domain Ontology O; maximum semantic distance d; minimum support min_sup;

Assumption: Web pages in WL are annotated with semantic information;

Output: (1) Frequent semantic objects;

(2) Semantic association rules;

Algorithm

1. // To find semantic association rich transaction
2. j=0;
3. SemW={}//Semantic Web Log
4. For i=1 to m do
5. While w_i contains semantic objects do
6. j=j+1;
7. SemW→SemW U {<o_j, t_i>}
8. End while
9. End for
10. For i=1 to j do
11. For k=1 to j do
12. $M_{Oi→}$|r| ∈R //number of edges to reach from o_i to o_j
13. End while
14. End for
15. Output f_0 = SemAwarePHS (M, SemW, d, min_sup)
16. Output semantic association rules by using f_0

Fig 2: Semantic Aware Framework

3.3.3 Semantic Aware PHS

SemAwarePHS (M, S, d, min_sup)

Input: sequence database S, Semantic distance Matrix M, Maximum semantic distance d, Minimum support min_sup,H_k: hash table for candidate k-item sets. L U T_k lookup table for storing coding relationship at k^{th} itemsets.

Output: Semantic rich frequent sequences

Algorithm:

1. Scan database S to find the set of frequent 1-sequence,L1={s_1,s_2,…s_n}
2. k=1;
3. C_1=L_1; //record frequent 1-sequnces where support>=min_sup;
4. k=2;
5. D_k=D; //get large item sets L_k (k>=2)
6. While(D_k!=Ø) do
7. D=D_k

//Construct Hk for candidate k-item sets where semantic distance of each pair is less than d

For all P,Q Є L_{k-1} with P={i_1,i_2,…i_{k-1}} and Q={i_1,i_2,…i'_{k-1}} and semdis (i_{k-1} , i'_{k-1}) <=d

For all 2-itemset x of t do

H_k .add(x);

// record the relationship if the item sets has support ≥min_sup

For each item sets y in H_k do

If H_k .count(y) then record the coding relationship on L U T_k ; //shrink database and get the next candidate item sets;

D_k =Ø

For each transaction t Є D do

Concatenate joinable pairs of 2-itemsets in t and then transform them into new 2-itemsets by looking them up in L U T_k

D_k U {t};

End

k++;

end

return decode(L U T_k)

Fig 3: Semantic Aware PHS

3.3.4 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. There are two algorithms namely Semantics-Aware Framework and Semantic-Aware PHS to generate the frequent semantic objects and semantic association rules. 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 domain side and minimum support count. And then the clean web logs are mapped with semantic objects and build the semantic matrix to calculate the 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 having frequent 1-itemsets, construct hash table for candidate two item sets 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 appearance and frequent 2-itemsets are generated. From next step to the final, the algorithm is different to former algorithms. The frequent 2-itemsets are unique corresponded to a word in a new system.

For example, candidate item set AB assign as unique word namely "a", and another BC assign as "b"(this may also called encoding). The correspondent of all sets must be stored in a table (called lookup table) and this table will be used to extract next frequent patterns. In this step, it also truncates the items which are not a frequent 2-itemsets from database. At this time, a new database is created and generates candidates' 3-itemsets in the old by joining two 2-itemsets or by join operator two words in new system. So, sets of 3-itemsets in the basic database will be in form of 2-itemsets under considering database. For example joining with a (AB) and b (BC) will result in (ab).

If these entire works have been finished, then the iteration end. By this processes go ahead, the database shrinks since all the non frequent items will be trimmed during the processing. So, the size of database in the next iteration is much smaller than in the previous. After finishing the repeat, we must retrace the previous steps to extract large item sets from the lookup tables. And then semantic association rules are generated to develop the recommendations. From theoretical point of view, it is better both on the side of complexity and running time since it does not only eliminate collision but also fix the length of candidate item sets to simplify the task of making hash function.

This proposed algorithm also prunes the transactions which do not contain any frequent items, and trims the non-frequent items from the transaction at each step. Moreover, this proposed algorithm performs better than the Apriori algorithm since at each step it reduces the database size to be scanned; it generates much smaller sized candidate 2-itemsets at the initial step.

3.3.5 Semantic Association Rules

Semantic association rules are rules that carry semantic information in them, and prediction system can get better informed decisions. Such rules are used to provide more accurate prediction than regular association rules by overcoming ambiguous predictions problem. For example, consider the following two semantic association rules: Such that $M_{o2,o5}<M_{o2,o4}$ meaning that o_5 is semantically closer to o_5 than o_4. Then the prediction result will prefer o_5 over o_4 and the pages representing product o_5 will be chosen. Such association rules can also be used for more intelligent user behavior analysis and the capability that is not provided by regular association rules.

3.3.6 Correlation based Similarity

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.

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

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 offline navigation paths.

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

$$R_{t,j} = \frac{P(t|j)}{\sum_{j=1,\dots n} P(t|j)}$$

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

$$P_{t,j} = \frac{freq(t\ in\ j)}{\sum_{j=1,\dots n} freq(t\ in\ j)}$$

4. SYSTEM OVERVIEW

In this proposed system, the web site structure is assigned as the basic framework to build reference ontology. The web server log file is the input of data preprocessing method. In this step, the log files which included the non-responded requests, requests by software agents and irrelevant requests are removed. And the user identification and session identification method are used in clean web logs. And then the user transaction files are described as the domain information and the domain ontology is build according to the related data from it. And then the PHS algorithm is joined with the domain ontology and the frequent item sets and semantic association rules are generated by using this algorithm.

In online phase, when the current user session is generated, the correlation based similarity is examined by applying the semantic association rules which were produced from ontology based PHS algorithm. And then the prediction list is produced for current user to retrieve information next.

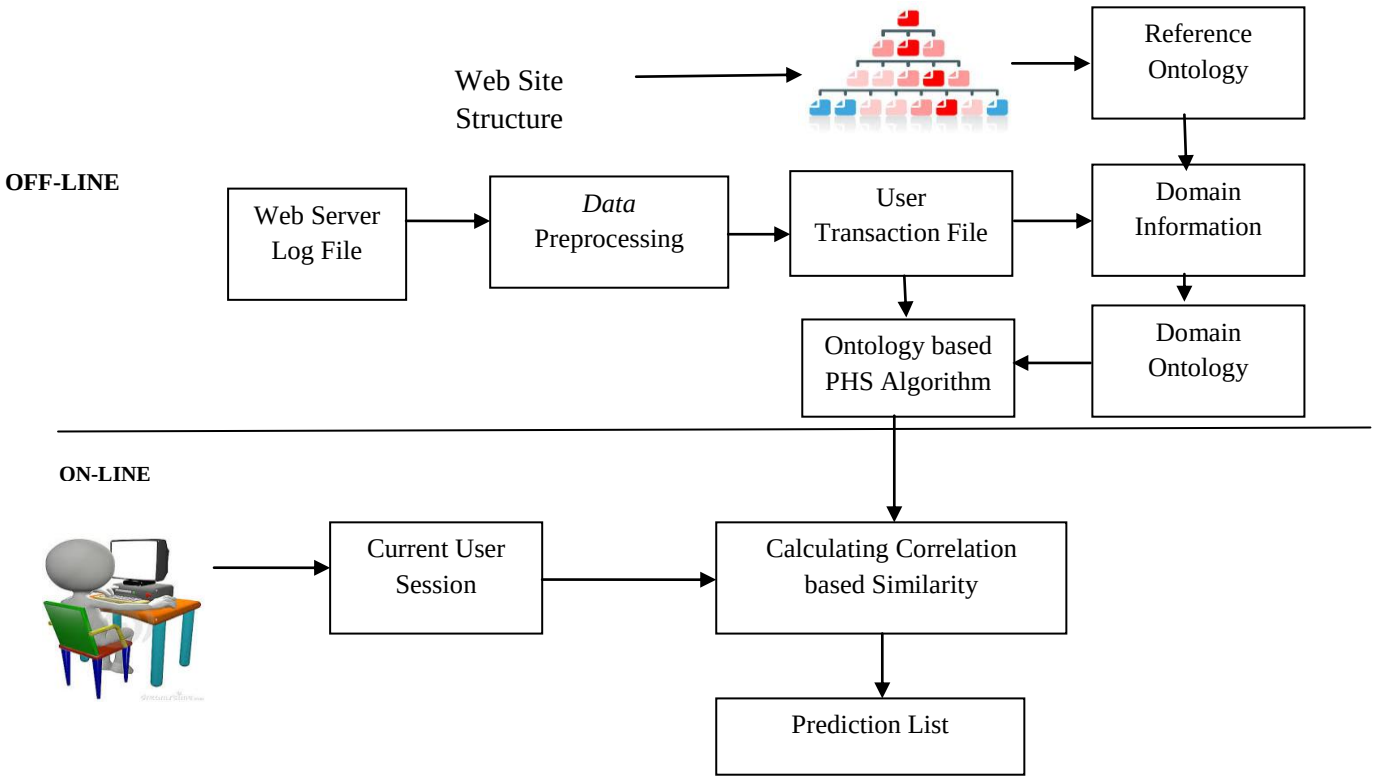

Fig 4: System Design

5. EVALUATION RESULT

In this proposed system, it is easy to understand advantages of PHS algorithm after studying Apriori algorithm and PHS algorithm. The main feature is that DHP (Direct Hash and Pruning) algorithm can reduce significantly the number of total scanning database by using hash mechanism. In addition, trimming redundant items and transactions during their process, so the database becomes smaller than smaller. Figure 5 shows that the difference of the execution time by using Apriori and DHP algorithm. DHP algorithm which is better performance than Apriori because it is not only reduced the number of scanning the whole database but also decrease the size of database.

But DHP has weakness such as the 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 sine in hash table and it may face

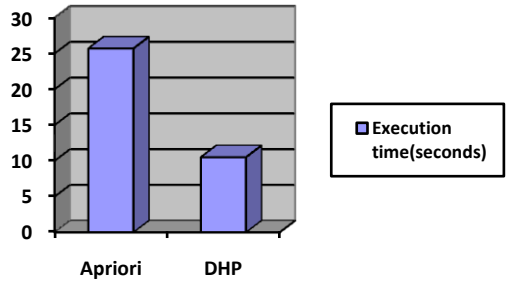

Fig 5: Execution time of Apriori and DHP

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.

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 6. 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

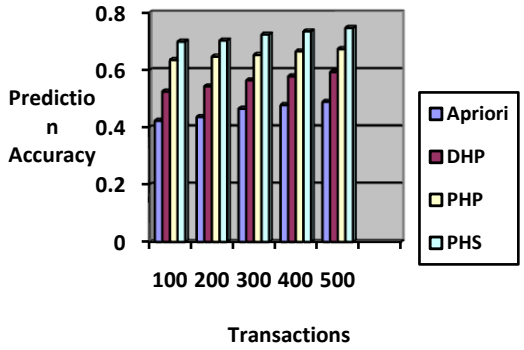

Fig 6: Prediction Accuracy Analysis- Apriori, DHP, PHP and PHS Algorithms

6. CONCLUSION

In this proposed system, the architecture of prediction system utilizes the ontology based PHS algorithm to extract the semantic association rules in offline phase and calculate the prediction list by using correlation based similarity to match with these rules and current user navigation paths. This proposed system overcomes the disadvantages of classical 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 proposed system has many steps to be processed than simple prediction system but can give more accurate results. 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. Compared with the conventional Web usage based prediction system, this proposed system can be extended to incorporate domain knowledge of the web application in the form of ontology in pattern-growth sequential pattern mining and other algorithms.

7. ACKNOWLEDGMENTS

This research paper is made possible through the help and support from everyone, including: parents, teachers, family, and friends and in essence, all sentient beings.

8. 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.

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

[6] 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.

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

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

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

[10] 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.