International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:8, No:1, 2014

Semantically Enriched Web Usage Mining for Personalization

Suresh Shirgave, Prakash Kulkarni, José Borges

Abstract**—**The continuous growth in the size of the World Wide Web has resulted in intricate Web sites, demanding enhanced user skills and more sophisticated tools to help the Web user to find the desired information. In order to make Web more user friendly, it is necessary to provide personalized services and recommendations to the Web user. For discovering interesting and frequent navigation patterns from Web server logs many Web usage mining techniques have been applied. The recommendation accuracy of usage based techniques can be improved by integrating Web site content and site structure in the personalization process.

Herein, we propose semantically enriched Web Usage Mining method for Personalization (SWUMP), an extension to solely usage based technique. This approach is a combination of the fields of Web Usage Mining and Semantic Web. In the proposed method, we envisage enriching the undirected graph derived from usage data with rich semantic information extracted from the Web pages and the Web site structure. The experimental results show that the SWUMP generates accurate recommendations and is able to achieve 10-20% better accuracy than the solely usage based model. The SWUMP addresses the new item problem inherent to solely usage based techniques.

Keywords**—**Prediction, Recommendation, Semantic Web Usage Mining, Web Usage Mining.

I. Introduction

THE number of Web pages available is increasing very rapidly adding to the hundreds of millions pages already on-line. The structure of Web sites is becoming more and more complex because of this rapid and in some cases chaotic growth. When searching and browsing a Web site, users are very often overwhelmed by the huge amount of information and are faced with the big challenge of finding the desired information in the right time. For the Web site owner the main issues that have to be dealt with are helping the users to find relevant information and providing personalization mechanisms to help them fulfill their information needs.

Web mining is a broad research area emerging to address the issues that arise due to the explosive growth of the Web and it is usually divided into three general categories: Web content mining, Web structure mining and Web usage mining. Web usage mining has been defined as the research field focused on developing techniques to model users' Web navigational behavior. According to [1], [2] most Web usage mining techniques that use solely usage data are based on association rules, sequential patterns and clustering. As noted in [3], usage based personalization has limitations in situations where there is insufficient usage data to extract patterns related to certain categories, when the site contents changes and when new pages are added but are not yet included in the Web log. To address these problems Web content and/or Web site structure can be incorporated with the usage data in order to improve the accuracy of the personalization process [4]. In the past many research efforts have tried to incorporate Web page content and Web site structure into the Web usage mining and personalization techniques, but very few have performed this using detailed semantic data inferred by means of Semantic Web Technologies.

In this work we propose to extend the WebPUM approach described in [5] with rich semantic data characterizing the contents of the web pages and Web site structure characterizing the topology of the web site. More precisely, we propose a semantically enriched Web Usage Mining method for Personalization (SWUMP) and argue that by incorporating semantic and structure data into WebPUM we will be able to improve the recommendation accuracy. We note that the WebPUM is based solely on usage data and it is not capable of capturing the information goals of the user. In addition, we expect the new method to be able to address new item problem.

WebPUM represents usage data by means of an adjacency matrix and induces the navigation patterns using a graph partitioning technique. Herein, we propose two methods to extend WebPUM. In the first method the adjacency matrix derived from usage data is enriched with the semantic data and the navigation patterns are induced. These navigation patterns are fed to recommendation engine. In the second method, the navigation patterns induced by the graph partitioning approach are augmented and augmented navigation patterns are enriched with the semantic metadata from the pages' contents. The navigation patterns are augmented to take into account Web site structure in the personalization process. In the augmentation process the pages that are linked to the pages in the navigation pattern and pages that are semantically similar, but not included in Web log are added in the navigation pattern. The semantic navigation patterns are then used to generate the recommendations.

The performance of the SWUMP is evaluated by means of extensive experiments conducted on both real world datasets (the Music Machine data set and the Semantic Web dog food Web site) and on a synthetically generated data set. The experimental results show that the recommendation accuracy

Suresh Shirgave is with the Textile and Engineering Institute, Ichalkaranji, Maharashtra, India-416115 (e-mail: skshirgave@yahoo.com).

Prakash Kulkarni is with the Walchand College of Engineering, Sangli, Maharashtra, India-416415 (e-mail: pjk_walchand@rediffmail.com).

José Borges is with the INESC TEC, Faculty of Engineering, University of Porto, R. Dr. Roberto Frias, 4200-465 Porto, Portugal (e-mail: jlborges@fe.up.pt).

of the SWUMP is superior to solely usage based method presented in [5].

In summary our key contributions in this paper are:

- The usage based approach WebPUM [5] is extended to take into account semantic metadata obtained from the page contents and Web site structure. The semantic metadata extracted takes into account both the semantics in a page contents and the semantic relationship in the Web pages.
- The new item problem of solely usage based technique is addressed by using a navigation pattern augmentation technique.
- Two recommendation algorithms that integrate content semantics and site structure with the users' navigational behavior are proposed.
- An extensive set of experiments which demonstrate the effectiveness of the proposed method was conducted.

The rest of this paper is organized as follows: In Section II, we review recent research advances in Web usage mining. In Section III, we briefly discuss WebPUM method which is the basis of our proposed method. Section IV describes the architecture of the proposed method. The overall performance of the proposed method is evaluated in Section V. Finally, Section VI provides concluding remarks and sheds light on future directions.

II. Related Work

Web usage mining techniques provide a complete process for the extraction of models from usage data. These models can be automatically exploited by a personalization system to generate recommendations. Many Web usage mining techniques integrate Web page content and site structure with usage data to improve accuracy of the recommendations.

A. Usage Based Techniques

Tak Yan et al. [6] proposed one of the first Web usage mining system. The method discovers clusters of users that exhibit similar information needs by examining user access logs. Based on which categories an individual user falls into, links are suggested dynamically to the user. The approach used for clustering is affected by several limitations related to scalability and the effectiveness of the results found. Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava [7] presented WebPersonalizer, a system that provides dynamic recommendations as a list of hypertext links to users. The method is based on anonymous usage data combined with the Web site structure. F. Masseglia, P. Poncelet, and R. Cicchetti [8] proposed an integrated system WebTool that relies on sequential patterns and association rules extraction to dynamically customize the hypertext organization. The current user's behavior is compared to one or more previously induced sequential patterns and navigational hints are provided to the user. Bamshed Mobasher, Robert Cooley, and Jaideep Srivastava [9] proposed an approach that captures common user profiles based on association rule discovery and usage-based clustering. The extracted knowledge is used to provide recommendations for users in real-time. The approach suggests visited pages, but is unable to include in the suggestions pages that were not visited by users. Ranieri Baraglia and Fabrizio Silvestri [10] proposed a Web usage mining system, SUGGEST, that is designed to dynamically generate personalized content of potential interest for users. Dimitrios Pierrakos et al. [11] proposed a method that exploits Web usage mining techniques in order to identify communities of Web users that exhibit similar navigational behavior with respect to a particular Web site. The information produced by the system can either be used by the administrator, in order to improve the structure of the Web site, or it can be fed directly to a personalization module to generate recommendations. B. Zhou, S. C. Hui, and K. Chang [12] proposed Sequential Web Access-based Recommender System (SWARS) that applies sequential access pattern mining to identify sequential Web access patterns with high frequencies. The Pattern-tree constructed from Web access patterns is used for matching and generating recommendations. José Borges and Mark Levene [13] presented a Variable Length Markov Chain (VLMC), which is an extension of a Markov chain that allows variable length history to be captured. The VLMC model has been shown to provide better prediction accuracy while controlling the number of states of the model.

B. Approaches Based On Usage and Content

Eirinaki et al. [14] presented a semantic Web personalization framework that combines usage data with Web contents (annotated in terms of ontology) in order to generate useful recommendations. Stuart Middleton, Nigel Shadbolt and David Roure [15] presented a recommender system for online academic publications where user profiling is done based on a research papers' topic ontology. Haibin Liu and Vlado Kešelj [16] proposed a novel approach for classifying navigation patterns and predicting users' future requests. The approach is based on the combined mining of Web server logs and the content of the Web pages represented in terms of character N-grams. The approach can be improved by using content representation technique that takes into account semantics of Web page contents. Xin Jin, Yanzan Zhou, and Bamshad Mobasher [17] proposed a unified framework which provides dynamic and personalized recommendations. The proposed framework is based on Probabilistic Latent Semantic Analysis to create models of Web users, taking into account both usage data and Web site contents. Miao Wan et al. [18] proposed a Random Indexing approach that is based on a vector space model, to discover intrinsic characteristics of Web users' activities. The Random Indexing with various weight functions is used for clustering individual navigational patterns and creating common user profiles. The clustering results will be used to predict and prefetch Web requests for grouped users. Pinar Senkul and Suleyman Salin [19] proposed a technique for integrating semantic information into Web navigation pattern generation process. The frequent navigational patterns are composed of ontology instances instead of Web page addresses and these are used for generating recommendations. Thi Thanh Sang Nguyen et al.

[20] proposed a novel ontology-style model of Web usage mining that enables the integration of Web usage data and domain knowledge to support semantic recommendations. The recommendations are generated by using Web user access sequences that are represented in Web Ontology Language (OWL). Juan D. Velásquez, Luis E. Dujovne, and Gaston L'Huillier [21] proposed a methodology for identifying Website Key Objects. Website Key Objects are the most appealing objects for users within a Website. The accurate extraction of Website Key Objects enables the possibility of enhancing the Web site by empowering the information that users are looking for. Mehdi Adda, Petko Valtchev, Rokia Missaoui [22] studied ontology based pattern space and proposed xPminer mining method. The xPminer performs a complete and non-redundant traversal of the pattern space and discovers all the frequent patterns. The mined frequent patterns are used to generate recommendations. Julia Hoxha, Martin Junghans, Sudhir Agarwal [23] presented an approach for the formalization of user Web browsing behavior across multiple sites. The usage logs are mapped to comprehensible events from the application domain. The semantic, formal description of each log is mapped to concepts of a vocabulary of the domain knowledge.

In summary, all of these works attempt to improve recommendation accuracy by integrating usage data, Web site structure and Web page contents. It is possible to generate more effective recommendations by incorporating detailed semantic data in the personalization process. The combined Web usage mining approaches, i.e. approaches that use usage data as well as Web page contents for personalization, can be extended by using detailed semantic metadata inferred from Web page contents and expressed by using semantic Web technology, RDF.

III. WEBPUM METHOD

The WebPUM approach presented in [5] is based solely on usage data. An undirected graph is induced from the user navigation sessions and a graph partitioning approach is applied to mine navigation patterns. In order to provide predictions the Longest Common Subsequence (LCS) algorithm is used to classify a user session into one of the navigation patterns and users' future movements are predicted as the pages in the pattern that are unseen by the user.

An undirected graph is constructed from the navigation sessions induced from Web server logs. In the process, an adjacency matrix is computed that represents degree of connectivity between the Web pages. The entry $W_{a,b}$ between page *a* and page *b* is calculated by using a time connectivity and a frequency measure. The Time connectivity measures the degree of visit ordering between two Web pages, and it is given by the formula,

$$TC_{a,b} = \frac{\sum_{i=1}^{N} \frac{T_i}{T_{ab}} \times \frac{f_a(k)}{f_b(k)}}{\sum_{i=1}^{N} \frac{T_i}{T_{ab}}} \quad (1)$$

where T_i is the total time duration of the i^{th} session that contain both the pages *a* and *b* and T_{ab} is difference between requested time of page *a* and page *b* in the session. The value of $f(k)$ is the position of the page in the session. The time connectivity measure is normalized to hold values between 0 and 1. The Frequency measures the co-occurrence of two pages in the sessions and it is given by,

$$FC_{a,b} = \frac{N_{ab}}{\max\{N_a, N_b\}} \quad (2)$$

where N_{ab} is the number of sessions containing both page *a* and *b*. N_a and N_b are number of session containing only page *a* and page *b*. The connectivity between any two pages is given by,

$$W_{a,b} = \frac{2 \times TC_{ab} \times FC_{ab}}{TC_{ab} + FC_{ab}} \quad (3)$$

Each entry $M_{a,b}$ of the adjacency matrix contains value of $W_{a,b}$ that represents the degree of connectivity between the two pages *a* and *b*. The undirected graph is created corresponding to the adjacency matrix. To limit the number of edges in the graph, if the value of $W_{a,b}$ is less than a threshold value (named as MinFreq) the edge is discarded. Further details on the undirected graph construction process from navigation sessions are available in [5].

For generating navigation patterns a graph partitioning algorithm is used. The graph partitioning algorithm finds the connected components in the undirected graph and it is based on Depth first search (DFS) algorithm. The vertices in a connected component represent a navigation pattern. The DFS algorithm is invoked repeatedly till all the vertices in the undirected graph are visited.

The LCS algorithm is used to classify the current active session into one of the navigation pattern and recommendations are generated. As described in [5] the WebPUM method does not takes into account other Web data like pages' content and the site structure.

IV. THE SWUMP METHOD

In this work we extend the WebPUM method proposed in [5] in order for it to incorporate site structure and page semantics in the personalization process to generate more precise recommendations. Fig. 1 illustrates the overall architecture of the proposed SWUMP method. The following subsections describe the components of the method in detail.

A. Web Log Preprocessing

As shown in Fig. 1, the pre-processing task is the first step in Web usage mining, being responsible for reading the Web logs and inducing the corresponding user navigation sessions.

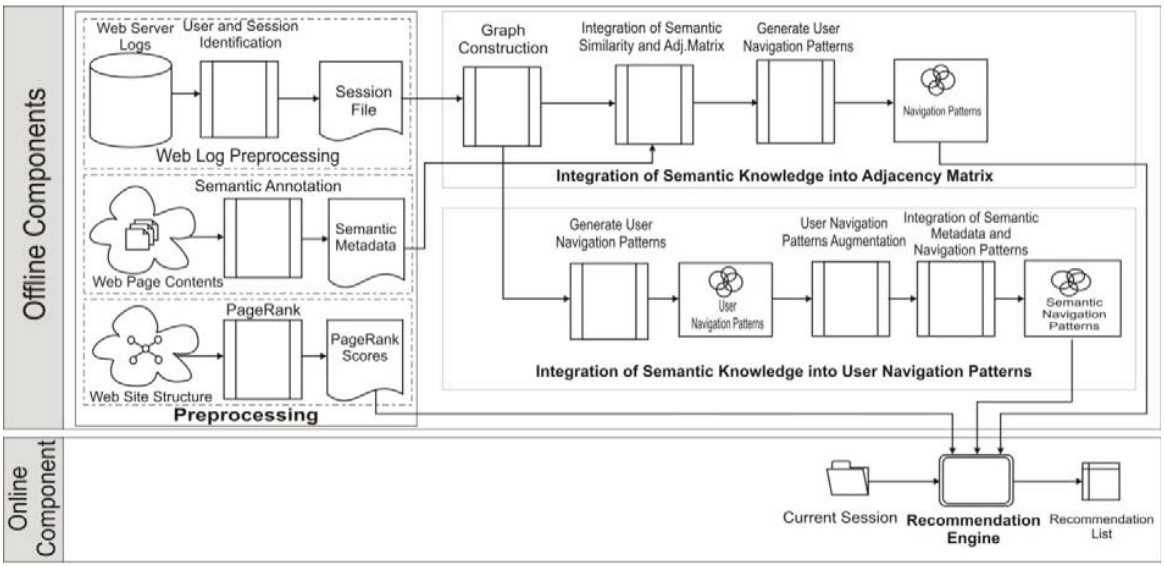

Fig. 1 The structure of the SWUMP

In the process, the log data is cleaned in order to remove entries that are not useful to represent user Web navigation behavior and for repairing erroneous data. Also, users are identified based on information available in the log file, such as the IP address, the type of operating system and the browsing software. The proposed method makes use of Web log pre-processing techniques described in [24].

B. Semantic Annotation

The Semantic Web provides a common framework that allows data to be shared and reused across applications, and enterprises, in a manner understandable by machines. Semantic annotation is a key component for the realization of the Semantic Web that formally identifies concepts and the relations between concepts in documents. The RDF is the standard data and modeling specification used to encode metadata and digital information.

The SWUMP makes use of the OpenCalais[1] and the AlchemyAPI[2] Web services for generating the semantic annotation of the Web pages, which includes topics, social tags, concept tags, keywords, search terms and other metadata.

The system crawls the Web site to collect the Web pages. The OpenCalais processes the pages and returns annotated semantic metadata as RDF payloads serialized as XML data containing the topics, social tags, identified entities, facts, and events. The metadata also contains the relations that involve at least one recognized entity from the content. Relations are generally all subject-predicate-object relationships without predefining their types. Web pages are also processed by the Alchemy API to generate complementary semantic metadata. AlchemyAPI utilizes statistical algorithms, natural language processing technology and machine learning algorithms to analyze Web page contents and extract keywords, search terms, concept tags, and information about people, places, companies, topics, languages and more. The AlchemyAPI has a concept tagging feature that automatically tags documents and text in a manner similar to human-based tagging. The results are also returned as RDF payloads.

The resulting XML data is parsed to extract the metadata and store it in the RDF data store. We make use of AllegroGraph RDF Data Store[3], which is a modern, high-performance, persistent RDF graph database. The semantic similarity between the Web pages is calculated using the method described in [25]. The method returns a similarity value between 0 and 1, where 1 means that the instances have exactly the same properties and 0 means no shared properties. In our method, if the similarity score is above certain threshold, then the Web pages are considered to be semantically similar. The semantic similarity between Web pages is represented in terms of a semantic similarity matrix. This similarity information is used to augment the navigation patterns as discussed in next subsection.

C. Navigation Patterns Augmentation

The navigation sessions extracted and clustered by using the graph partitioning algorithm are primarily based on the user sessions induced from the Web server logs. These sessions exclusively represent Web pages or resources that were visited by the user. It is possible that pages that are not included in the logs are relevant to the current active user session, and should be in the recommendation list generated. Such pages could be newly added pages or pages that have a link to them but are not presented to the user due to bad navigation design of the Web site. If the navigation patterns are generated only from Web server logs, the newly added Web pages cannot be recommended to the user.

To address this problem, in SWUMP the navigation

[1] http://www.opencalais.com
[2] http://www.alchemyapi.com
[3] http://www.franz.com/agraph/allegrograph

patterns are augmented to include the connected neighborhood of every page in a navigation pattern. The neighborhood of a page p is the set of all the pages directly linked from p, all the pages that directly link to p. The set of pages in the neighborhood of a page can be determined by using the Web site structure. In addition, each navigation patterns is augmented with the newly added Web pages that are highly semantically similar to the Web pages in each navigation pattern. We let $ANP = \{anp_1, anp_2,...,anp_k\}$, $1 \leq i \leq k$, be set of augmented navigation patterns, where k is the total number of navigation patterns. The augmentation step is applied only for generating semantic navigation patterns as discussed in next subsection.

D.Integration of Semantic Knowledge and Usage Data

The analysis of the contents of the user visited Web pages and other domain knowledge like Web site structure is a natural step to better capture users' information goals. Thus we enhance a solely usage based method WebPUM proposed in [5] and described in the Section III in order to take into account Web page contents and Web site structure. In this paper we have used two methods to integrate semantic metadata and usage data.

1. Integration of Semantic Knowledge into Adjacency Matrix

As discussed in Section III usage data is represented by using adjacency matrix M and $M_{i,j}$ is the value of $W_{i,j}$ between page i and j. For the integration of semantics into a usage data, we extend the approach presented in [26]. In [26] the semantic information used to characterize Web pages is obtained from a domain ontology that is provided by the ontology engineer during the design of the web site. The authors have assumed that a single web page represents a single concept from the ontology, which is not always the case in real world, and the semantic distance between two pages is calculated based on number of edges separating two pages in the domain ontology.

The SWUMP method makes use of semantic metadata to calculate the semantic similarity instead of distance in the domain ontology used in [26]. The semantic similarity is represented in terms of a semantic similarity matrix that gives the similarity score between every pair of Web pages. Thus, the semantic similarity matrix S is combined with the adjacency matrix M in order to derive the semantically enriched weight matrix T by using (4) as follows:

$$T_{p_i,p_j} = M_{p_i,p_j} + \begin{cases} \left(\frac{S_{p_i,p_j}}{\sum_{k=1}^{n} S_{p_i,p_k}}\right), S_{p_i,p_j} > 0 \\ 0 \quad , S_{p_i,p_j} = 0 \end{cases} \tag{4}$$

The semantic similarity matrix S is normalized in such a way that each entry is between 0 and 1. As defined in (4), to attain this normalization, each entry in the row is divided by row sum. In SWUMP a graph partitioning algorithm is applied on the semantically enriched matrix T in order to induce the navigation patterns. The set of navigation patterns generated are represented as $NP = \{np_1, np_2,...,np_k\}$, in which each np_i is a subset of the set of Web pages in the Web site. These navigation patterns are generated using semantically enriched matrix T. The semantic similarity between pages will have influence on the navigation patterns generated and lead to addition of new pages in the navigation pattern. This due to the fact that even though the connectivity weight between pages is zero (given by the usage data), there will be semantic similarity score value present in the combined matrix T. These navigation patterns are used for the next link of choice prediction and personalization process as discussed in next subsection.

2. Semantic Navigation Patterns

The navigation patterns are generated by applying the graph partitioning approach discussed in Section III before integrating semantic data into the adjacency matrix. The generated navigation patterns are augmented in order to address the new item problem, as discussed in previous subsection. Then, the augmented navigation pattern is integrated with the semantic metadata to create the semantic navigation patterns. Such integration makes use of an adaptation of the technique described in [16], which is focused on content represented by means of N-grams. Herein, we extend the technique in order to be able to make use of semantic metadata.

As a result, a semantic navigation pattern is composed of a collection of semantic metadata items that were extracted from the pages in the corresponding augmented navigation pattern. Each semantic navigation pattern is represented by n-dimensional vector v, where n is the total number of semantic metadata items from the complete set of Web pages in the site. The vector for each semantic augmented navigation pattern is given by, $v = \{w(f_1, p), w(f_2, p), \ldots, w(f_n, p)\}$, where $w(f_j, p)$, for $1 \leq j \leq n$, is the weight of the j^{th} semantic metadata item f_j in the pattern. Weight $w(f_j, p)$ is the normalized frequency of semantic metadata item f_j in the pattern and is given by,

$$w(f_j, p) = \frac{\text{Frequency}_{f_j,p}}{\sum_{f_j \in p} \text{Frequency}_{f_j,p}} \tag{5}$$

that corresponds to the frequency of item f_j in the pattern p divided by the sum of the frequencies of all items in the pattern p.

In addition, for each semantic metadata item f_j in the pattern we define its document frequency $df(f_j, p)$ as the number of Web pages in the pattern in which the item f_j occurs. Document frequency is given by,

$$df(f_j, p) = \frac{n_{f_j}}{N_p} \tag{6}$$

where n_{fj} is the number of documents containing f_j and N_p the total number of pages. Since $df(f_j,p)$ has been normalized it will have value between 0 and 1. A high value of $df(f_j,p)$ indicates that the item carries more representative information in the pattern.

Let DF be the threshold value of document frequency. By varying the DF it is possible to configure different semantic

navigation patterns and filter out semantic items that are not very frequent. A semantic navigation pattern does not include the semantic metadata items whose document frequency is below *DF*. Based on the performance comparison of semantic navigation patterns with different document frequency values, we will be able to find out which value of document frequency will generate more accurate recommendations. The set of semantic navigation patterns is represented as, $PS = \{v_1, v_2, \ldots, v_k\}$, where k is the number of patterns. The semantic metadata based navigation patterns associate the meaning of the Web pages' content with the user navigation patterns. These semantically enriched navigation patterns that better captures information goals of a Web user are used to generate recommendations as discussed next in subsection.

E. Recommendation Engine

As stated in [27], "Web recommendation is a promising technology that attempts to predict the interests of Web users, by providing the users information and/or services that they need without the users explicitly asking for them". The recommendation engine is the online component of a recommendation system.

1. Recommendations Using Navigation Pattern Generated from Combined Matrix

As discussed in previous subsection navigation patterns are generated by applying the graph partitioning approach on the semantically enriched adjacency matrix. The generated navigation patterns, $NP = \{np_1, np_2, \ldots, np_k\}$, are used to generate recommendations. The Longest Common Subsequences (LCS) [29] algorithm is utilized to classify the current active user session into one of the navigation pattern np_i that is highly similar to current active user session. The recommendation set is generated from the set of Web pages in the navigation pattern np_i. If the number of Web pages in the generated recommendation set is more than the size of recommendation set N, then Web pages in the generated recommendation set will be arranged in the order according to decreasing values of PageRank score [28] and only the top N pages are added in the recommendation set while the rest of Web pages are not considered in the recommendation set.

2. Recommendations Using Semantic Navigation Patterns

The current active user session is classified into the semantically enriched navigation pattern, PS_i, where $i=1,2,\ldots,k$, that better describes the user information goals. Recommendations are then generated from the set of pages in the augmented navigation pattern anp_i. The current active user session is vectorized following the procedure introduced in previous subsection. In order to classify the active user session p into the adequate semantic navigation pattern PS_i, a dissimilarity metric $D(p, PS_i)$, $i=1,2,\ldots,k$, is calculated to assess the similarity between the current active user session and each of the patterns. The user session is associated to the pattern corresponding to the smallest value of $D(p, PS_i)$.

The accuracy of this classification depends on the clever choice of dissimilarity measure. The dissimilarity measure takes two profiles as an input and returns a positive number and for the two identical profiles, the dissimilarity is 0. In our proposed method, we use dissimilarity proposed in [30].

$$d_3(\mathrm{tf_p}, \mathrm{tf_{PS_i}}) = \sum_{x_i \,\epsilon\, \text{profile}} \frac{|\,\mathrm{tf_p}(x_i) - \mathrm{tf_{PS_i}}(x_i)|}{\sqrt{\mathrm{tf_p}(x_i) \times \mathrm{tf_{PS_i}}(x_i)} + 1} \qquad (7)$$

The current active user session is classified into one of the navigation pattern PS_i and recommendations are generated from the set of pages in the augmented navigation pattern anp_i. To limit the number of pages in the recommendation set PageRank score is used.

V. Experimental Evaluation

In this section we provide a detailed experimental evaluation of the proposed method SWUMP. In next subsections we state the datasets description, evaluation metrics, and experimental results and its discussion.

A. Data Sets Description

For the experimental evaluation of the SWUMP approach it is necessary that datasets provide both the server log data and the Web page contents. These experiments have been conducted on the publicly available Music Machine data set (DS-1)[4], on the Semantic Web dog food Web site (DS-2)[5], and on a synthetic usage data generated for a university Web site (DS-3). The DS-1 data set is provided cleaned and sessionized, and we have used access entries in a four month period, from January to April 1999. For DS-2 we have used the access entries from June 2010 to December 2010 for the Semantic Web Dog Food Web site. This is a very active Web site of publications, people and organizations in the Web and semantic Web fields, covering several of the major conferences and workshops. Finally, the DS-3 corresponds to a Web site of a technical university including Web pages of individuals (i.e. students and teachers), news group and courses, for which the usage data was generated using a technique similar to the described in [31].

Table I depicts summary statistics of the experimental data sets. For each data set, we indicate the total number of access entries, the number of clean access entries (that is obtained after removing entries that are not useful to represent user Web navigation behavior), the number of pages occurring in the log, the total pages identified by crawler during crawling of the Web site and the total users identified. We also give the total number of sessions derived from each data set and the number of sessions of lengths more than two; session length is measured by the number of requests a session is composed of. We assume that the induced user sessions that have a length of more than two pages are more suitable for the experiments since it might carry more information about Web users' intention on the Web site. Therefore, sessions having less than three page requests were filtered out from the datasets.

[4] http://machines.hyperreal.org
[5] http://data.semanticweb.org

TABLE I
STATISTICS OF EXPERIMENTAL DATA SET

Attributes	DS-1	DS-2	DS-3
Total access entries	936677	452192	1325198
Clean access entries	936677	430252	1325198
Total Web page accessed in log	850	1919	835
Total pages Identified by Crawler	1037	2105	1050
Different access users	116183	23245	50000
Total Identified sessions	143633	26667	50000
Total Identified sessions (≥2 requests)	91926	21067	49040

B. Evaluation Metrics

In order to evaluate the effectiveness of the recommendations generated by the proposed method the performance is determined using three different standard measures, namely precision, coverage, and the F1 measure [32]. Among these, precision and coverage metrics have been widely used in recommender system research. As precision and coverage are inversely related, a combination measure, called the F1 measure, is used to give equal weight to both precision and coverage [32]. While, precision measures the degree to which the recommendation engine produces accurate recommendations, coverage measures the ability of the recommendation engine to recommend all the pages that are likely to be visited by the user.

C. Experimental Results

A series of experiments focused on evaluating the performance of the SWUMP as compared to the purely usage based WebPUM technique [5] were conducted. Cross validation with k = 5 subsets was used, being the sessions split k subsets, the model is built from $k - 1$ subsets, leaving the k^{th} subset as a test set. In order to simulate active sessions of a Web user, each test session is split into two parts. The first part of the session simulates an active session of the current user and the second part the Web pages that the user will request during his further navigation on the Web site. That is, the first part of the session is used to predict its second part. Each active session is then fed into the recommendation engine in order to produce a recommendation set. The recommendation set obtained is then compared to the second part of the test session in order to compute the precision, coverage, and F1 measure metrics [32].

Experiments were conducted to assess recommendation methods proposed in previous subsection. For the experimentation we have chosen recommendation set size as twelve. By increasing recommendation set size it is likely that coverage can be improved, but reduces the precision. It is observed during experimentation that F1 measure is maximized for recommendation set size of twelve, indicating that the best balance of precision and recall is achieved for typical recommendation set sizes. The performance of the SWUMP can be tuned by varying the value of MinFreq from zero to one.

Table II summarizes the average values of Precision, coverage and F1 measure for the WebPUM approach [5] (M_1), the proposed recommendation method based on navigation patterns derived from semantically enriched adjacency matrix (M_2) and the recommendation method that makes used of semantic navigation patterns (M_3). The results obtained for the MinFreq values of 0.0, 0.1, and 1.0 are not significant and hence not reported in the Table II.

As shown in Table II, the results obtained for the proposed SWUMP method shows more accurate values for precision, coverage and F1 measure in comparison to the solely usage based technique WebPUM. The WebPUM and proposed method achieve better performance for the MinFreq range 0.4 to 0.6. Both the recommendation methods proposed in this paper outperform the WebPUM approach. The recommendation method that use navigation patterns derived from semantically enriched adjacency matrix achieves 10-15% performance improvement over the WebPUM and recommendation method that use semantic navigation patterns achieves 15-20% performance improvement over the WebPUM method. It is clear from figures in Table II that the recommendation method that uses semantic navigation patterns outperforms recommendation method that use navigation patterns derived from semantically enriched adjacency matrix. The accuracy of recommendations generated by using navigation patterns derived from semantically enriched adjacency matrix indicates that clusters generated are compact and integration of semantics in the adjacency matrix improves accuracy of the clustering. The WebPUM approach achieved the best results when we choose the value of MinFreq in the range 0.5 to 0.6 and in case of proposed method 0.4 to 0.6. The results show that the proposed method is able to improve the accuracy of recommendations for different values of MinFreq.

The experimental results reveal that the usage, content and Web site structure information together improves the recommendation quality. It can be observed that recommendations generated by SWUMP are better than those obtained by WebPUM model for all the three data sets. The experimental results indicates that our approach for generating recommendations by integrating usage, content and structure is able to improve the accuracy of recommendations in the personalization process.

VI. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we proposed Semantically enriched Web Usage Mining method for Personalization (SWUMP), an extension of usage based method WebPUM. The SWUMP is used to predict users' future requests by combining usage data, Web site structure and detailed semantic information extracted from Web page contents. Two recommendation methods are proposed. The first generates recommendations using navigation patterns derived from semantically enriched adjacency matrix, and second generates recommendations using semantic navigation patterns.

Results of extensive experimental evaluation conducted on three data sets are reported. The experimental results show that incorporating semantic data and site structure into WebPUM method improves recommendation accuracy. The semantic Web mining that combines semantic Web and Web usage mining, results in a more accurate classification of

navigation patterns, and leads to a more accurate prediction of users' future requests and accurate recommendations as compared to pure usage based techniques. It is observed that recommendation method that makes use of semantic navigation patterns outperforms the method using navigation patterns derived from semantically enriched adjacency matrix.

TABLE II
RESULTS OF RECOMMENDATION ENGINE FOR THREE DATA SETS DS-1, DS-2, AND DS-3

MinFreq	Precision			Coverage			F1 Measure		
	M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3
Data Set : Music Machine (DS-1)									
0.2	0.05	0.06	0.06	0.09	0.12	0.12	0.06	0.08	0.08
0.3	0.17	0.47	**0.50**	0.21	**0.51**	**0.56**	0.18	0.49	**0.53**
0.4	0.40	**0.56**	**0.61**	0.45	**0.63**	**0.68**	0.42	**0.59**	**0.64**
0.5	0.46	**0.59**	**0.64**	0.49	**0.63**	**0.68**	0.47	**0.60**	**0.66**
0.6	0.47	**0.60**	**0.65**	0.46	**0.59**	**0.64**	0.46	**0.59**	**0.64**
0.7	0.25	0.32	0.34	0.22	0.28	0.30	0.23	0.29	0.32
0.8	0.15	0.19	0.20	0.12	0.15	0.16	0.13	0.17	0.18
0.9	0.10	0.13	0.13	0.09	0.12	0.12	0.09	0.12	0.13
Data Set : Semantic Web Dog Food Web site (DS-2)									
0.2	0.10	0.12	0.13	0.18	0.22	0.24	0.12	0.15	0.17
0.3	0.22	**0.56**	**0.60**	0.27	**0.62**	**0.66**	0.24	**0.58**	**0.63**
0.4	0.43	**0.59**	**0.63**	0.48	**0.64**	**0.69**	0.45	**0.61**	**0.65**
0.5	0.47	**0.57**	**0.61**	**0.52**	**0.64**	**0.68**	0.49	**0.60**	**0.65**
0.6	0.49	**0.60**	**0.64**	**0.50**	**0.62**	**0.65**	0.49	**0.60**	**0.65**
0.7	0.21	0.26	0.27	0.28	0.34	0.31	0.24	0.29	0.31
0.8	0.15	0.18	0.19	0.19	0.23	0.21	0.16	0.20	0.21
0.9	0.08	0.10	0.10	0.10	0.12	0.11	0.08	0.11	0.11
Data Set : Synthetic (DS-3)									
0.2	0.11	0.14	0.14	0.18	0.22	0.66	0.13	0.16	0.18
0.3	0.25	**0.60**	**0.63**	0.27	**0.62**	**0.67**	0.25	**0.61**	**0.65**
0.4	0.47	**0.62**	**0.66**	0.49	**0.63**	**0.67**	0.47	**0.62**	**0.67**
0.5	0.49	**0.60**	**0.64**	**0.52**	**0.64**	**0.67**	0.50	**0.62**	**0.66**
0.6	0.49	**0.59**	**0.63**	**0.50**	**0.61**	**0.64**	0.49	**0.60**	**0.64**
0.7	0.20	0.25	0.26	0.26	0.32	0.33	0.22	0.27	0.29
0.8	0.15	0.18	0.19	0.19	0.24	0.25	0.16	0.20	0.22
0.9	0.09	0.11	0.12	0.09	0.11	0.11	0.09	0.11	0.11

There are some aspects in which the proposed method can be improved. Due to dynamic nature of the Web, researchers have recently paid more attention to mining evolving Web user profiles that vary with time. The proposed method can also be extended for a database backed Web site that generates the Web pages dynamically based on structured queries performed against backend databases. The contents of Web page depends on query parameters, hence these parameters must be taken into account in the personalization process.

REFERENCES

[1] Sungjune Park, Nallan Suresh, and Bong-Keun Jeong, "Sequence-based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-means algorithm," Data & Knowledge Engineering, vol. 65, pp. 512–543, 2008.

[2] Bing Liu, Web Data Mining, Second Edition ed.: Springer, 2011.

[3] Bamshad Mobasher, Hoghua Dai, Tao Luo, Yuqing Sun, and Jiang Zhu, "Integrating Web Usage and Content Mining for More Effective Personalization," in Proceedings of the International Conference on E-Commerce and Web Technologies, Greenwich, UK, 2000.

[4] Magdalini Eirinaki, Michalis Vazirgiannis, and Iraklis Varlamis, "Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process," in Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'03), Washington DC, 2003.

[5] Mehrdad Jalali, Norwati Mustapha, Md. Nasir Sulaiman, and Ali Mamat, "WebPUM: A Web-based recommendation system to predict user future movements," Expert Systems with Applications, vol. 37, pp. 6201-6212, 2010.

[6] Tak Woon Yan, Matthew Jacobsen, Hector Garcia-Molina, and Umeshwar Dayal, "From User Access Patterns to Dynamic Hypertext Linking," Computer Networks and ISDN Systems, vol. 28, no. (7–11), pp. 1007–1014, 1996.

[7] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava, "Automatic personalization based on Web usage mining," Communications of the ACM, vol. 43, no. 8, pp. 142–151, 2000.

[8] F. Masseglia, P. Poncelet, and R. Cicchetti, "WebTool: An Integrated Framework for Data Mining," in Proceedings of the 9th International Conference on Database and Expert Systems Applications (DEXA'99), Florence, Italy, 1999, pp. 892-901.

[9] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava, "Creating Adaptive Web Sites Through Usage-Based Clustering of URLs," in Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99), November 1999.

[10] Ranieri Baraglia and Fabrizio Silvestri, "An online recommender system for large Web sites," in Proceedings of the IEEE/WIC/ACM international conference on Web, Beijing, China, 2004.

[11] Dimitrios Pierrakos, Georgios Paliouras, Christos Papatheodorou, and Constantine D. Spyropoulos, "KOINOTITES:A Web Usage Mining Tool for Personalization," in Proceedings of the Panhellenic Conference on Human Computer Interaction, 2001.

[12] B. Zhou, S. C. Hui, and K. Chang, "An intelligent recommender system using sequential Web access patterns," in IEEE conference on cybernetics and intelligent systems, 2004, pp. 393–398.

[13] José Borges and Mark Levene, "Evaluating Variable Length Markov Chain Models for Analysis of User Web Navigation Sessions," IEEE Trans. on Knowledge And Data Engineering, vol. 19, no. 4, pp. 441 – 452, Apr 2007.
[14] Magdalini Eirinaki, Dimitrios Mavroeidis, George Tsatsaronis, and Michalis Vazirgiannis, "Introducing Semantics in Web Personalization: The Role of Ontologies," in Proc. EWMF/KDO'2005, 2005, pp. 147-162.
[15] Stuart Middleton, Nigel Shadbolt, and David Roure, "Ontological User Profiling in Recommender Systems," ACM Transactions on Information Systems, vol. 22, no. 1, pp. 54–88, 2004.
[16] Haibin Liu and Vlado Kešelj, "Combined mining of Web server logs and Web contents for classifying user navigation patterns and predicting users' future requests," Data & Knowledge Engineering, vol. 61, no. 2, pp. 304–330, 2007.
[17] Xin Jin, Yanzan Zhou, and Bamshad Mobasher, "A Unified Approach to Personalization Based on Probabilistic Latent Semantic Models of Web Usage and Content," in AAAI Workshop on Semantic Web Personalization (SWP'04), July 2004.
[18] Miao Wan, Arne Jönsson, Cong Wang, and Lixiang Li, "Web user clustering and Web prefetching using Random Indexing with weight functions," Knowl Information Systems, October 2011.
[19] Pinar Senkul and Suleyman Salin, "Improving pattern quality in web usage mining by using semantic information," Knowledge and Information Systems, p. 2011.
[20] Thi Thanh Sang Nguyen, Hai Yan Lu, and Jie Lu, "Ontology-Style Web Usage Model for Semantic Web Applications," in 10th Int'l Conference on Intelligent Systems Design and Applications (ISDA), 2010, pp. 784-789.
[21] Juan D. Velásquez, Luis E. Dujovne, and Gaston L'Huillier, "Extracting significant Website Key Objects: A Semantic Web mining approach mining approach ," Engineering Applications of Artificial Intelligence, vol. 24, pp. 1532-1541, March 2011.
[22] Mehdi Adda, Petko Valtchev, and Rokia Missaoui, "A framework for mining meaningful usage patterns within a semantically enhanced web portal," in Proceedings of the Third C* Conference on Computer Science and Software Engineering C3S2E '10, New York, USA, 2010, pp. 138-147.
[23] Julia Hoxha, Martin Junghans, and Sudhir Agarwal, "Enabling Semantic Analysis of User Browsing Patterns in the Web of Data," in Julia Hoxha, Martin Junghans, Sudhir Agarwal, Lyon, France, 2012.
[24] R. Cooley, B. Mobasher, and J. Srivastava, "Data preparation for mining world wide web browsing patterns," Knowledge and Information System, vol. 1, pp. 5–32, 1999.
[25] Gunnar Grimnes, Peter Edwards, and Alun Preece, "Instance Based Clustering of Semantic Web Resources," in Proceedings of the 5th European Semantic Web Conference, LNCS Springer-Verlag, 2008.
[26] N. R Mabroukeh and C. I. Ezeife, "Semantic-rich Markov Models for Web Prefetching," in IEEE International Conference on Data Mining Workshops, 2009, pp. 465-470.
[27] G. Castellano, A. M. Fanelli, and M. A. Torsello, "NEWER: A system for NEuro-fuzzy WEb Recommendation," Applied Soft Computing, vol. 11, no. 1, pp. 793-806, January 2011.
[28] Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," Computer Networks, vol. 30, no. 1-7, pp. 107-117, 1998.
[29] Alberto Apostolico, "String editing and longest common subsequences," in Handbook of Formal Languages., 1997, pp. 361–398.
[30] Andrija Tomovic, Predrag Janicic, and Vlado Kešelj, "N-gram-based classification and hierarchical clustering of genome sequences," Computer Methods and Programs in Biomedicine, 2005.
[31] Peter I. Hofgesang and Jan Peter Patist, "On Modelling and Synthetically Generating Web Usage Data," in Int'l Conference on Web Intelligence and Intelligent Agent Technology, 2008, pp. 98-102.
[32] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval.: Addison Wesley, 1999.