International Journal of Web & Semantic Technology (IJWesT) Vol.4, No.4, October 2013

SEMANTICALLY ENRICHED WEB USAGE MINING FOR PREDICTING USER FUTURE MOVEMENTS

Suresh Shirgave[1] and Prakash Kulkarni[2]

[1] Department of Computer Science and Engineering, Textile and Engineering Institute, Ichalkaranji, India
[2] Department of Computer Science and Engineering, Walchand College of Engineering, Sangli, India

ABSTRACT

Explosive and quick growth of the World Wide Web has resulted in intricate Web sites, demanding enhanced user skills and sophisticated tools to help the Web user to find the desired information. Finding desired information on the Web has become a critical ingredient of everyday personal, educational, and business life. Thus, there is a demand for more sophisticated tools to help the user to navigate a Web site and find the desired information. The users must be provided with information and services specific to their needs, rather than an undifferentiated mass of information. For discovering interesting and frequent navigation patterns from Web server logs many Web usage mining techniques have been applied. The recommendation accuracy of solely 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 (SWUM), which combines the fields of Web Usage Mining and Semantic Web. In the proposed method, the undirected graph derived from usage data is enriched with rich semantic information extracted from the Web pages and the Web site structure. The experimental results show that the SWUM generates accurate recommendations with integration of usage, semantic data and Web site structure. The results shows that proposed method is able to achieve 10-20% better accuracy than the solely usage based model, and 5-8% better than an ontology based model.

KEYWORDS

Prediction, Recommendation, Semantic Web Usage Mining, Web Usage Mining

1. INTRODUCTION

The World Wide Web has become the biggest and the most popular way of communicating, retrieving and disseminating information. The number of Web pages available is increasing very rapidly adding to the hundreds of millions pages already on-line. The rapid and chaotic growth has resulted into more complex structure of Web sites. When searching and browsing a Web site, users are often overwhelmed by 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. Often, this results in

DOI : 10.5121/ijwest.2013.4405

higher visitor retention, increased profits for online store owners, in addition to helping users in finding the desired information. Thus, automated tools focused on helping users to search, extract, and filter the desired information and resources are very useful [1].

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 content mining is focused on the development of techniques to assist users in finding Web documents that meet a certain criteria. Web structure mining analyses the hyperlink structure of Web and it usually involves analysis of in-links and out-links of Web pages to, for example, rank search engine results. 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]. Many research efforts incorporate Web page content and Web site structure with Web usage mining and personalization techniques, but not many have used emerging semantic Web technologies and detailed semantic data in the process.

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 (SWUM) and argue that by incorporating semantic and site 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 a 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. 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. The performance of the SWUM 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 of the SWUM is superior to solely usage based method presented in [5] and combined mining method [6] that makes use of ontology to represent Web page contents.

In summary our key contributions in this paper are:

- The solely 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.
- A recommendation algorithm that integrates content semantics and site structure with the users' navigational behavior is proposed.
- An extensive set of experiments which demonstrate the effectiveness of the proposed method.

The structure of the paper is organized as follows: In Section 2, we review recent research advances in Web usage mining. In Section 3, we briefly discuss WebPUM method which is the basis of our proposed method. Section 4 describes the architecture of the proposed method. The overall performance of the proposed method is evaluated in Section 5. Finally, Section 6 provides concluding remarks and sheds light on future directions.

2. RELATED WORK

Several models have been proposed for modelling user browsing behaviour on a Web site and generating recommendations for a Web user. 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.

2.1. Usage Based Techniques

Tak Yan *et al.* [7] 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 *et al.* [8] 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 *et al.* [9] proposed an integrated system, WebTool, that relies on sequential patterns and association rules extraction to dynamically customize the hypertext organization. The current user's behaviour is compared to one or more previously induced sequential patterns and navigational hints are provided to the user. Ranieri Baraglia *et al.* [10] proposed a Web usage mining system, SUGGEST, that is designed to dynamically generate personalized content of potential interest for users. Bamshed Mobasher *et al.* [11] 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. Dimitrios Pierrakos *et al*. [12] proposed a method that exploits Web usage mining techniques in order to identify communities of Web users that exhibit similar navigational behaviour 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 *et al.* [13] 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 *et al.* [14] presented a Variable Length Markov Chain (VLMC) model, 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.

2.2. Approaches based on Usage and Content

Eirinaki *et al*. [6] 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 *et al.* [15] presented a recommender system for online academic publications where user profiling is done based on a research papers' topic ontology. Haibin Liu *et al.* [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 *et al.* [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 *et al.* [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).

2.3. Other Approaches

Juan D. Velásquez *et al.* [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 *et al.* [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 *et al.* [23] presented an approach for the formalization of user Web browsing behaviour 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. A. C. M. Fong *et al.* [24] proposed a semantic Web usage mining approach for discovering periodic Web access patterns from annotated Web usage logs. This approach highlights fuzzy logic to represent real-life temporal concepts and requested resource attributes of periodic pattern-based Web access activities.

2.4. Summary and Discussion

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.

3. WEBPUM METHOD

The WebPUM approach presented in [5] is based solely on usage data. 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 $W_{a,b}$ in the adjacency matrix 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,

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

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,

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

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,

$$W_{a,b} = \frac{2 \text{ x } TC_{ab} \text{ x } FC_{ab}}{TC_{ab} + FC_{ab}} \tag{3}$$

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.

4. SWUM METHOD

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

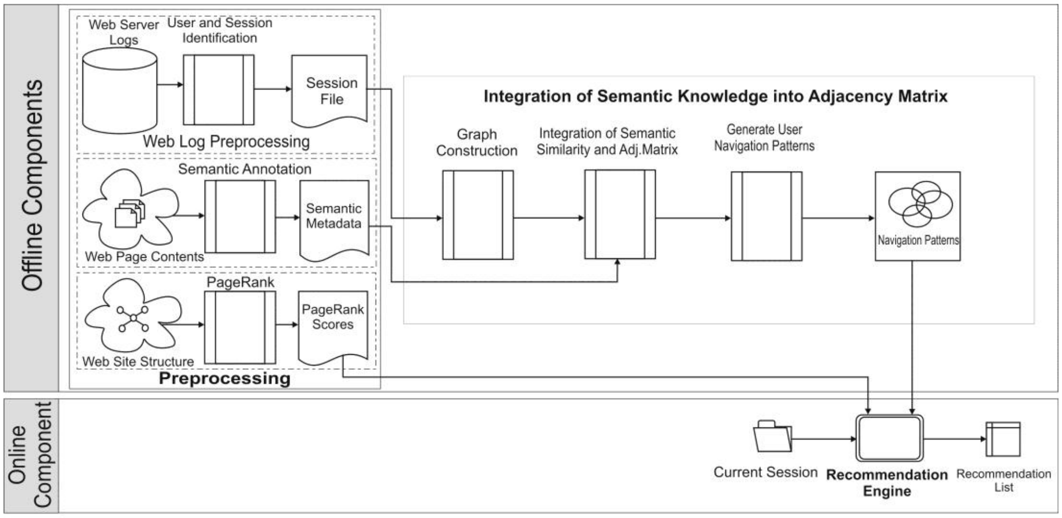

Figure 1 The structure of the SWUM

4.1. Web Log Preprocessing

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. In the process, the log data is cleaned in order to remove entries that are not useful to model the user Web navigation behaviour and for repairing erroneous data. User identification is based on information available in the log file, such as the IP address, the type of operating system and the browsing software. User navigation sessions are derived from the log file. The sessionization task consists of grouping a sequence of users' page requests into a unit named session. A session can be defined as an ordered collection of pages accessed by a user in a time window defined by the moment he entered the site and the moment he left it. The proposed method makes use of Web log pre-processing techniques described in [25].

4.2. 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 modelling specification used to encode metadata and digital information.

The SWUM 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

[1] http://www.opencalais.com
[2] http://www.alchemyapi.com

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 metadata information is used to calculate the semantic similarity between Web pages. The semantic similarity between the Web pages is calculated using the method described in [26]. 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. The semantic similarity between Web pages is represented in terms of a semantic similarity matrix and used to integrate in the adjacency matrix as discussed in next Section.

4.3. Integration of Semantic Knowledge into Adjacency Matrix

Web users' information goals are better captured by a detailed analysis of the contents of the user visited Web pages and other domain knowledge like Web site structure. Thus we enhance a solely usage based method WebPUM proposed in [5] and described in the Section 3 in order to take into account Web page contents and Web site structure. In this paper we have integrated semantic metadata, usage data and site structure in the personalization process to generate more accurate recommendations. As discussed in Section 3 usage data is represented by using adjacency matrix M and $M_{\mathrm{i,j}}$ is the value of $W_{i,\mathrm{j}}$ between page i and j. For the integration of semantics into a usage data, we extend the approach presented in [27]. In [27] 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 SWUM method makes use of semantic metadata to calculate the semantic similarity instead of distance in the domain ontology used in [27]. 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 Eq. (4) as follows:

$$T_{p_i,p_j} = M_{p_i,p_j} + S_{p_i,p_j} \tag{4}$$

The matrix T is obtained by combining usage data and semantic similarity between Web pages. In SWUM 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

[3] http://www.franz.com/agraph/allegrograph

= {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 Section 4.4.

4.4. Recommendation Engine

As stated in [28], "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. As discussed in Section 4.3 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,...,np_k}, are used to generate recommendations. The current active user session is classified into one of the navigation pattern np_i that is highly similar to current active user session using Longest Common Subsequences (LCS) [29] algorithm. 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 [30] 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.

5. EXPERIMENTAL EVALUATION

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

5.1. Data Sets Description

For the experimental evaluation of the SWUM 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 provided is 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].

[4] http://machines.hyperreal.org

[5] http://data.semanticweb.org

Table 1 depicts summary statistics of the experimental data sets. For each data set, we indicate the total number of access entries, number of clean access entries (that is obtained after removing entries that are not useful to represent user Web navigation behaviour), number of pages occurring in the log, 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.

Table 1 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

Figure 2 shows the distribution of session length for the three data sets. For example, session length of two indicates the percentage of sessions with two page requests that occur in the collection of sessions. As shown in Figure 2, the percentage of total sessions decreases when session length increases.

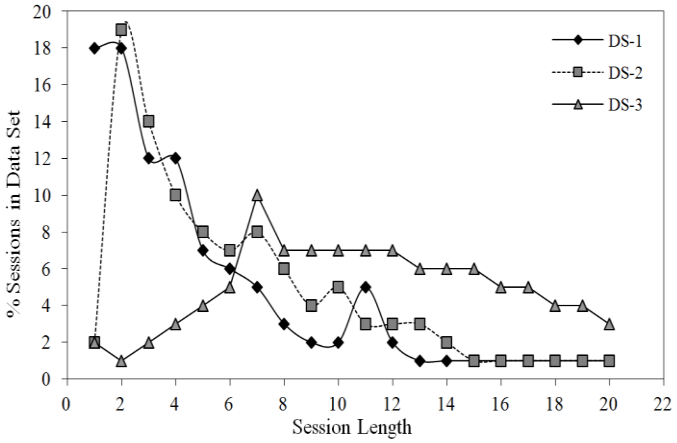

Figure 2 Sessions distribution of the data sets

5.2. 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.

5.3. Experimental Results

A series of experiments focused on evaluating the performance of the SWUM as compared to the solely usage based WebPUM technique [5] and ontology based method [6] 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 method proposed in Section 4.4. 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 SWUM can be tuned by varying the value of MinFreq from zero to one.

Table 2 summarizes the average values of Precision, coverage and F1 measure for the WebPUM approach [5] (M_1), ontology based method [6] (M_2) and the proposed recommendation method based on navigation patterns derived from semantically enriched adjacency matrix (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 2. The concept of MinFreq is not applicable to ontology based method $M_{2.}$ As shown in Table 2, the results obtained for the proposed SWUM method shows more accurate values for precision, coverage and F1 measure in comparison to the solely usage based technique WebPUM and ontology based method. The WebPUM and proposed method achieve better performance for the MinFreq range 0.4 to 0.6. The recommendation method proposed in this paper outperforms the WebPUM approach and ontology based method. The proposed recommendation method achieves 10-20% performance improvement over the WebPUM and 5-8% performance improvement over ontology method. 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 increase in the accuracy of the proposed method over ontology based method clearly indicates advantage of using detailed semantic metadata instead of ontology in the personalization process.

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

The experimental results reveal that the usage, content and Web site structure information together improves the recommendation accuracy. It can be observed that recommendations generated by SWUM are better than those obtained by WebPUM model and ontology based method 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.

6. CONCLUSIONS AND FUTURE DIRECTIONS

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

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 solely usage based techniques and ontology based method.

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

[2] 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.

[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] 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.

[7] 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.

[8] 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.

[9] 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.

[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] 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.

[12] 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.

[13] 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.

[14] 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.

[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] A.C. M. Fong, Baoyao Zhou, Siu C. Hui, Jie Tang, and Guan Y. Hong, "Generation of personalized ontology based on consumer emotion and behavior analysis," IEEE Transactions on Affective Computing, 2011.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] Alberto Apostolico, "String editing and longest common subsequences ," in Handbook of Formal Languages., 1997, pp. 361–398.

[30] 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.

[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.

Authors

Suresh Shirgave is currently Associate Professor of Computer Science and Engineering at DKTE Society's Textile and Engineering Institute, Ichalkaranji, Maharashatra, India. He received his bachelors and masters degrees in Computer Science and Engineering from Walchand College of Engineering, Sangli, Maharashatra, India. He is currently pursuing Ph.D. in Computer Science and Engineering from Shivaji University, Kolhapur, India. He has published many research papers in national and international conferences and Journals. His research interests include data mining, Web usage mining, recommendation, social networks and Internet security.

Prakash Kulkarni completed his master’s and Ph. D. Research studies in Electronics Engineering from Shivaji University, Kolhapur in 1986 and 1993 respectively. He is currently working as a professor in Computer Science and Engineering department of Walchand College of Engineering, Sangli. He has provided guidance to many PhD students in the areas of Electronics Engineering and Computer Science and Engineering. He has executed many research proposals of AICTE, New Delhi. He has published 11 International and 10 national journal papers. His research interest includes Computer Vision, Pattern recognition, Artificial Neural Networks, Data Mining, Web mining and Information retrieval. He received a distinguished position in Asia Pacific’s “Who’s Who 2004” reference dictionary. He is also a recipient of Best Teacher Award of Maharashtra State Government for the year 2011–2012.