International Journal of Information Technology & Decision Making
Vol. 13, No. 4 (2014) 721–753
© World Scientific Publishing Company
DOI: 10.1142/S0219622014500643

World Scientific
www.worldscientific.com

Semantically Enriched Variable Length Markov Chain Model for Analysis of User Web Navigation Sessions

Suresh Shirgave*

Department of Computer Science and Engineering
Textile and Engineering Institute, Rajwada
Ichalkaranji, Maharashatra 416115, India
skshirgave@yahoo.com

Prakash Kulkarni

Department of Computer Science and Engineering
Walchand College of Engineering, Vishrambag
Sangli, Maharashatra 416115, India
pjk_walchand@rediffmail.com

José Borges

INESC TEC, Faculty of Engineering
University of Porto, R. Dr. Roberto Frias
Porto 4200-465, Portugal
jlborges@fe.up.pt

Published 22 May 2014

The rapid growth of the World Wide Web has resulted in intricate Web sites, demanding enhanced user skills to find the required information and more sophisticated tools that are able to generate apt recommendations. Markov Chains have been widely used to generate next-page recommendations; however, accuracy of such models is limited. Herein, we propose the novel Semantic Variable Length Markov Chain Model (SVLMC) that combines the fields of Web Usage Mining and Semantic Web by enriching the Markov transition probability matrix with rich semantic information extracted from Web pages. We show that the method is able to enhance the prediction accuracy relatively to usage-based higher order Markov models and to semantic higher order Markov models based on ontology of concepts. In addition, the proposed model is able to handle the problem of ambiguous predictions. An extensive experimental evaluation was conducted on two real-world data sets and on one partially generated data set. The results show that the proposed model is able to achieve 15–20% better accuracy than the usage-based Markov model, 8–15% better than the semantic ontology Markov model and 7–12% better than semantic-pruned Selective Markov Model. In summary, the SVLMC is the first work proposing the integration of a rich set of detailed semantic information into higher order Web usage Markov models and experimental results reveal that the inclusion of detailed semantic data enhances the prediction ability of Markov models.

Keywords: Web usage mining; Markov chain models; recommendation; prediction; semantic web usage mining.

*Corresponding author.

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 keep growing very rapidly, adding to hundreds of millions pages already online. 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 research discipline that uses content, structure, and usage statistics in the process of helping to turn the World Wide Web into a more useful resource, by facilitating search for the desired information. 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 navigation data. Modeling user's navigational behavior on a Web site is useful in applications such as Web caching, Web page recommendation, Web search engine results ranking and Web page personalization.[2,3]

According to Ref. 4, Web usage mining techniques are mainly based on stochastic Markov models, association rules, sequential patterns and clustering techniques. Markov models have been widely used for predicting user's next link of choice, for predicting longer sequences of navigation options, for pre-fetching links and in adaptive Web site applications. Up to now, only a few research efforts have tried to incorporate Web page content into Web usage mining and personalization process, and even fewer have performed this using Semantic Web Technologies. To the best of our knowledge, works reporting the integration of usage data and semantic data rely solely on semantics, expressed in ontology terms. Such ontologies reflect the main concepts of a Web site and usually each page is annotated with the semantic concepts manually at the time of Web site creation.[5,6] These works assume that the domain knowledge in the form of domain ontology is available at the time of Web site design.

In this work, we propose a method that combines a higher order Markov model with rich semantic data automatically extracted from Web sites' page content. We propose a new Semantic Variable Length Markov Chain Model (SVLMC) that combines a higher order Markov model[7] and detailed semantic data that characterizes the contents of Web pages, which is obtained by means of two open source Web services.

We believe that using semantic data with higher level of detail is a step forward from ontology-based models. As a result, it will be possible to match closer the users' topics of interest while navigating and, therefore, to improve recommendation accuracy. In SVLMC, the semantic metadata is automatically extracted from Web pages' contents and it consists, for example, the main entities, facts, events, social tags and topics. A recommendation engine uses usage data (a higher order Markov model) integrated with this semantic data (in a Resource Description Framework (RDF) data store) to generate recommendations to a user.

Previous works have characterized Web page contents simply by a set of keywords representing the main concepts in combination with either Markov chain models or other usage-based models. Keywords are used to compute similarity between Web pages, based on exact matching between these keywords. In these techniques, only a binary matching between documents is achieved, whereas no actual semantic similarity is taken into consideration. Hence keyword-based approaches are incapable of capturing semantic information present in Web pages.

In some of the ontology-based methods, each page is allowed to represent a single concept in the ontology.[6] These methods are not able to capture semantics present in a Web page and are, therefore, incapable of capturing the semantic level relationships among Web pages. This may lead to inaccurate representation of navigation behavior of a user. In the method proposed in Ref. 6, the semantic similarity between two pages is calculated from the distance in the ontology provided by the ontology engineer during the design of a Web site. This method assumes that each Web page represents a single concept and may lead to inaccurate value of similarity between Web pages.

In the method proposed in this paper, detailed semantic metadata is used to calculate the similarity between Web pages, leading to a more accurate representation of semantic similarity between Web pages. We believe that a more accurate semantic similarity between pages will lead to more accurate recommendations. To the best of our knowledge, in the context of Web usage mining, this is the first work combining a higher order Variable Length Markov Chain (VLMC) model with detailed semantic data induced directly from the Web pages' contents.

We study two methods to provide next pages' predictions and we evaluate the performance of the SVLMC by means of extensive experiments conducted on both real-world data sets (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 SVLMC model is superior to both the solely usage-based Markov model and to the other models that use simpler semantic data, providing better prediction and recommendation accuracy.

The structure of the paper is organized as follows: Sec. 2 describes the related work in Web Usage Mining. Section 3 describes the system design and VLMC model which is the basis of our proposed model. Section 4 presents an experimental evaluation of the proposed model, and Sec. 5 provides concluding remarks and future directions.

2. Related Work

Several models have been proposed for modeling user browsing behavior on a Web site. Markov chain-based models are adequate for the context because they are compact, easy to understand, expressive and based on well-established theory. While some Web Usage Mining tools do not take into account the sequence of page views in a user session, Markov chain models are able, for example, to predict the next link of choice based on the sequence of pages previously visited by the user to reach the current page.[8] Numerous papers have dealt with methods to tackle prediction and recommendation problem using data mining techniques and Markov models.

2.1. *Approaches based on data mining techniques*

Many data mining techniques have been applied for extracting useful patterns from Web logs.[9] Oh and Kim[10] proposed a new similarity measure to compute similarity between two sequences and developed a hierarchical clustering algorithm. This clustering algorithm is able to generate better-quality clusters than traditional clustering algorithms. These clusters can be used to generate recommendations. Huang[11] proposed an approach to mine association rules and sequential patterns from a large collection of Web log data by means of interestingness measures and pruning methods. Effectiveness in finding interesting association rules and sequential patterns is evaluated by using a number of interestingness measures. Park *et al.*[12] proposed a general sequence-based clustering for Web usage mining using K-means algorithm with artificial neural networks. The sequence-based clustering method generates accurate user clusters. Makris *et al.*[13] proposed a method to cluster similar user sessions into groups and build a generalized weighted suffix tree to represent each cluster. This method generates recommendations by returning all pages on the outgoing edges of the matching suffix. Jalali *et al.*[14] presented a framework for online prediction using a Web usage mining recommendation system and proposed a novel approach named WebPUM to classify navigation patterns for predicting users' future requests. In order to generate navigation patterns, an undirected graph based on connectivity between each pair of Web pages is constructed. Carmona *et al.*[15] proposed a method for extracting useful information from an e-commerce Web site. This method makes use of unsupervised and supervised descriptive data mining algorithms to obtain rules that help a webmaster team to improve the design of a Web site. Guerbas *et al.*[16] proposed an efficient framework for Web log mining and online navigational behavior prediction. The density-based clustering algorithm namely DBSCAN is used to mine navigational patterns. An inverted index built from all identified sessions from the log is used for generating recommendations.

2.2. *Markov models for prediction in Web usage mining*

Jespersen *et al.*[17] studied the quality of a fixed-order Markov model in representing a collection of navigation sessions and concluded that higher accuracy cannot be achieved with a fixed-order Markov model. Sarukkai[18] proposed path analysis using Markov Chains for representing user traversals in the Web space. Levene and Loizou[19] first introduced the concept of state cloning in the context of usage mining, where a method is proposed that applies the cloning operation to states whose first- and second-order probabilities diverge, by duplicating such states in a way that separates their in-links. The method aims at accurately representing second-order probabilities, which take into account the history of the Web page a user visited prior to clicking on a link leading to another page. Borges and Levene[7,8] have presented a VLMC model to extend a first-order Markov model in a way that is able to incorporate higher order probabilities. The VLMC model has been shown to provide better prediction accuracy while controlling the number of states of the model. Borges and Levene[20] studied scoring metrics for evaluating the prediction accuracy of methods for solving the prediction problem. These metrics can be used for comparing prediction methods.

2.3. *Enhanced Markov models for Web usage mining*

Ching *et al.*[21] proposed a higher order Markov chain model for categorical data sequences and applied it to the server logs data. K-means clustering algorithm is used to cluster user sessions derived from the Web log files and Web page clusters are used to construct a higher order Markov chain model. Their tests based on a realistic Web logs show improvement in the prediction accuracy. This approach integrates semantics' sequences and Markov chains in a context in which the focus is the analysis of sequences of semantic events.

Another work in a similar context was presented by Leonardi *et al.*[22] proposing a method suitable for salient event detection in soccer using audio and visual information. The focus of the paper is goal detection in soccer games and makes use of semantic characterization of a multimedia content, and the goal is the design of a semantic indexing system. Again, sequences of semantic events are analyzed and Markov chains are used in the process.

In the context of Web usage mining, Khalil *et al.*[23] combined lower order All-Kth Markov models with association rules techniques in order to give more predictive power for a Markov model while retaining small space complexity. Chimphlee *et al.*[24] introduced a Web access prediction model that integrates rough set clustering with a Markov model. However, this method has a major drawback, the lack of prediction accuracy. Deshpande and Karypis[25] proposed Selective Markov model (SMM), which stores only some of the states within the model. SMM intelligently selects parts of different order Markov model and generates a model which has reduced state complexity and high prediction accuracy. This scheme is not suitable for very large data sets. Eirinaki *et al.*[26] proposed a hybrid probabilistic predictive model to address several shortcomings of Markov predictive models by means of combining link analysis and Web usage mining techniques.

2.4. *Semantically enriched Markov models for Web usage mining*

Eirinaki *et al.*[27] presented a Semantic Web Personalization framework that combines usage data with Web contents (annotated in terms of ontology) in order to generate useful recommendations. Zhang and Nasraoui[28] proposed a novel method to improve the recommendation accuracy of a Markov model by combining the Markov model with traditional content-based search techniques. This method makes use of first- and second-order Markov model to represent clickstream data. To compensate for the sparsity of the Markov model and improve coverage, a search engine is used to add similar content pages in the recommendation set. The limitation of this approach is that traditional content-based search does not take into account semantics. Zhang *et al.*[29] proposed a semantic session analysis method for partitioning Web usage logs. This method enhances usage logs with semantic information. A Markov chain model based on ontology semantic measurement is used to classify active session and predict users' future navigation. Mabroukeh and Ezeife[5] proposed a scheme to use semantic information as the criteria for pruning states in a higher order SMM. Maximum semantic distance has been used as a measure for pruning higher order Markov models. The semantic rich Markov model when compared with higher order SMM has smaller size and provides nearly equal accuracy. Mabroukeh and Ezeife[6] presented an approach to integrate semantic information directly in the transition probability matrix of lower order Markov models, resulting in semantic-rich lower order Markov models. This integration is able to achieve less space complexity and accurate prediction with lower order Markov Models. We believe that model can be enhanced by gathering richer semantic information rather than simple semantic distance in the domain ontology provided by the ontology engineer during the design of a Web site.

2.5. *Semantically enriched Web usage mining techniques*

Hu and Zhong[30] extended Web mining to Web farming and proposed a unified model for integrating multiple Web information sources. This model analyzes customer behavior to actively influence a customer's decision making. The customers' surfing and purchasing behavior is tied together by using clickstream logs collected at the application layer. The existing commercial Web sites and portals can use this model as a common plug-in. Fong *et al.*[31] 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. Nguyen *et al.*[32] proposed a novel ontology-style model of Web usage mining that enables conjugation of Web usage data and domain knowledge to support semantic Web recommendations to a Web user. A user Web access sequences are represented in Web Ontology Language (OWL) and used for generating the recommendations. Senkul and Salin[33] 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.

2.6. *Summary of discussion*

In summary, the works referred in the preceding sections attempt to improve recommendation accuracy by applying data mining techniques, Markov process and by integrating usage data, Web site structure and Web page contents. The works that have used Web site content in the personalization process have characterized Web page contents simply by a set of keywords, by N-grams or by ontology concepts. Such methods of characterization do not take into account the semantics in Web pages' content being unable to include in the recommendation set pages having semantically relevant content. We argue that it is possible to generate more effective recommendations in the context of Web usage mining by incorporating detailed semantic data extracted from Web page contents in the personalization process. The combined Web usage mining approaches, i.e., approaches that use usage data (modeled by using Markov chain model) as well as Web page contents for personalization, can be extended by using detailed semantic metadata inferred from Web page contents.

3. System Design

In this work, we propose the SVLMC that integrates usage data and Web page semantic contents. Figure 1 illustrates the overall architecture of the proposed system, which consists of five major modules: Web log Preprocessing, VLMC Model, Web Content Modeling, Semantic Rich Markov Model, and Recommendation Engine. The following subsections describe each component of the system in detail.

3.1. *Web log preprocessing*

The first module of the system is responsible for preprocessing tasks. The preprocessing task is an essential step in Web Usage Mining, thus the Web-log preprocessing

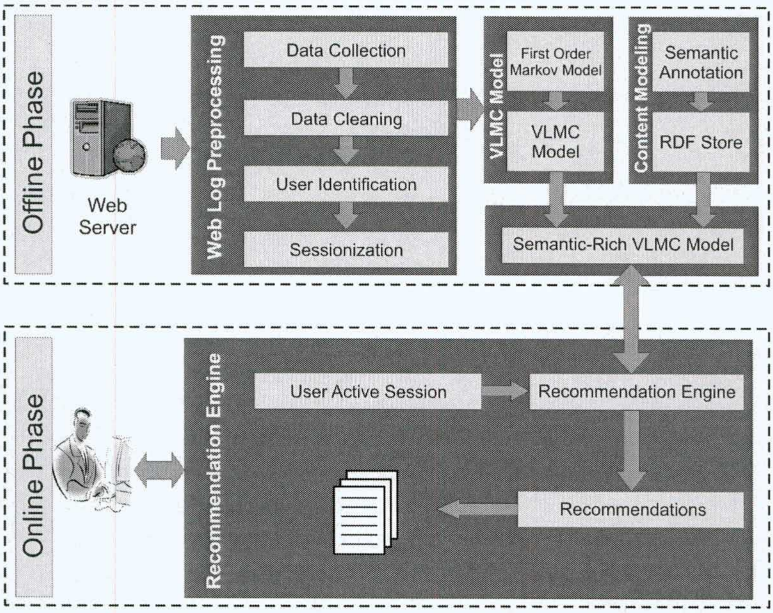

Fig. 1. Architecture of SVLMC.

component is responsible for reading original Web logs and identify the corresponding user Web navigation sessions.

3.1.1. *Data collection*

The primary data sources used in Web usage mining are the Web server log files. Web servers usually register the users' access activities on the Web site as Web server logs. Due to different Web server characteristics and parameter settings, there are many types of Web logs, but typically a log file captures information such as: client IP address, request time, requested URL, HTTP status code, and referrer. Each hit against the server corresponds to a HTTP request and generates a single entry in the server access logs.

3.1.2. *Data cleaning*

Not all the entries in Web server logs are useful for Web usage mining. The data cleaning task is responsible for removing entries that are not useful to represent user Web navigation behavior and for repairing erroneous data. In this process, requests that do not reflect human navigation behavior, i.e., requests for image files, JavaScript files (JS), page style files (CSS), requests from robots, automated tools and other software agents, have to be removed. Erroneous requests are removed from the data set.

3.1.3. *User identification*

User Identification is a task of associating a page request to a user who accessed the corresponding Web page. In general, the IP address of the machine placing a request is stored in the server log. The user identification task is not trivial due to existence of local caches, corporate firewalls, and proxy servers. In fact, it is not always possible to uniquely assign a Web page request to an individual user. Many user identification mechanisms have been developed and applied in Web Usage Mining research. Since users are treated as anonymous in most Web servers, two heuristic strategies, the proactive strategy and the reactive strategy, have been proposed to help differentiate the users.[34] In our system, we apply the reactive strategy to Web logs, and approximate users in terms of IP address, type of operating system and browsing software.

3.1.4. *Sessionization*

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. This session identification task takes all of the page references from a given user and splits them into a collection of unit sessions. For this process, two time-oriented heuristic methods, session-duration-based and page-stay-time-based, have been specifically proposed by Mobasher *et al.*[34–36] In the proposed SVLMC system both heuristics have been implemented.

3.2. *Variable length Markov chain model*

A VLMC is a stochastic model, generated by extending the notion of a first-order Markov chain by allowing a variable length history of page link choices to be captured within the model.[37] The VLMC addresses problem of inaccuracy of the first-order Markov model, and of the high state space complexity of higher order n-gram Markov models by modeling higher order dependencies between states of the Markov chain only in cases when the conditional probabilities are statistically significantly different.

We now give definition of probability of a session. Let $T = s_1, s_2, \dots, s_m$ be a sequence of pages visited by a user in the course of a navigation sessions. Using the chain rule, the probability estimate of T is given by,[37]

$$P(T) = P(s_1) \prod_{i=2}^{m} P(s_i \mid (s_1 \dots s_{i-1}). \tag{1}$$

In practice, full information of the conditional probabilities on the right-hand side of Eq. (1) is not available, and therefore estimation must be done with knowledge available from past activities of user's navigating through the site. So, the estimate of the conditional probability is given by,[37]

$$P(s_i \mid s_1, \dots, s_{i-1}) \approx P_M(s_i \mid s_{i-d_i}, \dots, s_{i-1}) = \frac{P_M(s_{i-d_i}, \dots, s_{i-1}, s_i)}{P_M(s_{i-d_i}, \dots, s_{i-1})}, \tag{2}$$

where s_{i-di} is the maximal suffix of the trail $s_1, s_2, \dots, s_i$, d_i represents order of the Markov chain and P_M represents the transition probability distribution of the d_i-order Markov chain. Let $C(T)$ denote frequency count of T, then estimate of conditional probability in Eq. (2) using the frequency counts can be denoted as,[37]

$$P_M(s_i \mid s_{i-d_i} s_2, \dots, s_{i-1}) \approx \frac{C(s_{i-d_i}, \dots, s_{i-1}, s_i)}{C(s_{i-d_i}, \dots, s_{i-1})}, \tag{3}$$

where, as before, d_i is the largest order of the Markov chain. The VLMC model extends a first-order Markov model by incorporating second-order (or higher order) probabilities. The first step in constructing VLMC model is building a first-order model.

3.2.1. *First-order Markov model*

In a first-order Markov model, there is a state corresponding to each Web page and a link connecting every two pages viewed in sequence. For each state that corresponds to a Web page, the page identifier is given. The ratio of the number of times a page was viewed to the total number of page views is a probability estimate for a user choosing the corresponding page from a set of all pages in a site. For each link, the proportion of times it was followed after viewing the anchor page gives the probability estimate of the two pages being viewed in sequence.

Table 1. A collection of Web pages in fictitious Website.

Request	ID
/index.html	1
/about_institute/about.html	2
/academic/academics.html	3
/academic/programs.html	4
/admissions/policy.html	5
/departments/it.html	6
/research/research.html	7

Table 2. Collection of sessions.

Session	Frequency
S,1,3,5,F	4
S,1,3,6,F	3
S,1,4,6,F	2
S,1,4,7,F	3
S,1,5,F	1
S,2,4,F	3
S,2,7,F	2
S,2,4,6,F	2
S,2,4,7,F	3

We will now give an example to illustrate the first-order model. Table 1 gives a collection of Web pages, and Table 2 gives a collection of sessions within the corresponding site. Figure 2 represents the first-order model induced from the sessions given in Table 2. Next to a transition, the first number indicates the number of times the corresponding hypertext link was traversed and the number in parentheses represents the estimated transition probability. For example, the probability of a transition from page 3 to page 5 is $p_{3,5} = 0.57$.

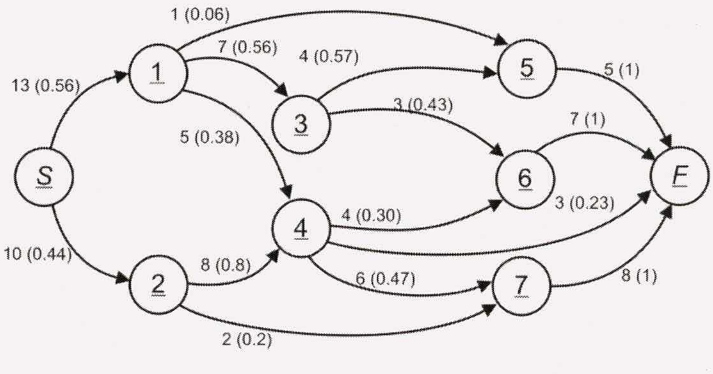

Fig. 2. First-order model.

3.2.2. *VLMC model*

Borges and Levene[7,8] proposed to use the cloning concept and a dynamic clustering-based method to improve the model accuracy in representing a collection of sessions. The method incorporates a clustering technique which determines an efficient way to distribute in-links among states. A second-order transition probability, $p_{i,k,j}$, gives probability of the transition (A_k, A_j) given that the previous transition that occurred was (A_i, A_k) and is computed as $p_{i,k,j} = w_{i,k,j}/w_{i,k}$, where $w_{i,k,j}$ and $w_{k,j}$ stands for corresponding 3-gram and 2-gram counts.[8] For example, according to the sessions in Table 2 we have that $p_{1,4,7} = 3/6 = 0.5$. A state needs to be cloned based on a link if it is not accurate, that is, if the absolute difference between the link's first- and second-order probabilities is greater than a specified accuracy threshold parameter, γ. The model is accurate if every state is accurate. In the example shown in Fig. 3, we have that $|p_{1,4} - p_{1,4,F}| = |0.23 - 0.0| = 0.23$. Therefore for any value of accuracy threshold below <0.23, state A_4 is not accurate and, therefore, the model is not accurate. The application of the cloning operation to state A_4 enables to increase the model accuracy and, when A_4 is cloned, state $A_{4'}$ is created. Since state A_4 has two in-links each of the in-links is assigned to one of the two states. K-means clustering algorithm is applied to the collection of second-order vectors, in order to identify groups of similar vectors with respect to γ. In this example, no clustering is needed since the cloned state had only two in-links, each of the in-links is assigned to one of the states.

3.3. *Content modeling-semantic annotation*

The SVLMC system employs semantic Web technology tools to represent contents of a Web site in order to extend the model with machine-readable metadata. RDF is a

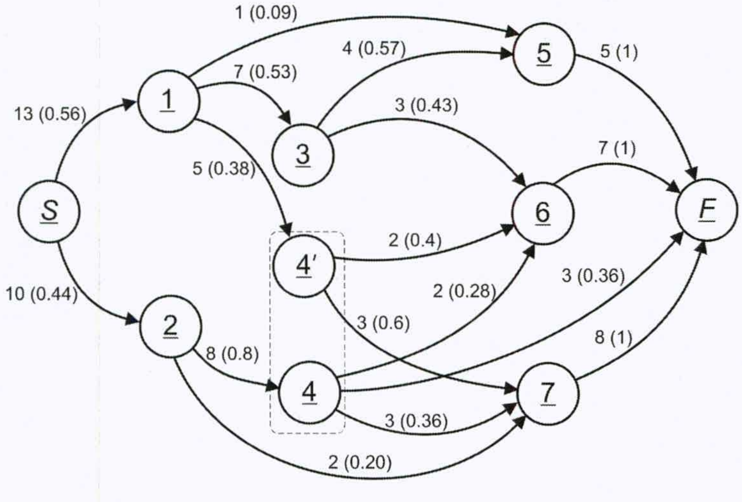

Fig. 3. The second-order Markov model.

standard data and modeling specification used to encode metadata and digital information.

Semantic annotation is a key component for the realization of the Semantic Web that formally identifies concepts and the relations between concepts in the documents. The proposed SVLMC system makes use of the OpenCalais[a] and the AlchemyAPI[b] Web services for generating the semantic annotation of the Web pages. Although there are many semantic annotation tools available, we have chosen OpenCalais and AlchemyAPI because these tools analyze the text and return rich semantic metadata automatically by applying the state-of-the-art machine learning, text mining and Natural Language Processing (NLP) algorithms. They have powerful Text categorization functionality, good recognition results and support for multiple languages (English, French and Spanish). These Web services return semantic metadata in a machine processable format (RDF) that provide the foundation for richer analytical and inference services. These inference services can uncover previously unknown relationships between resources. OpenCalais and AlchemyAPI have been widely used for Search Engine Optimization (SEO), improving search and navigation within a Web site by the semantic applications, and personalization mechanisms. These Web services have fast response time and are easily scalable. Hence they are suitable for Web sites having large number of pages. The SVLMC has a Web crawler component that downloads Web pages from the selected Web site. It takes a specified URL as a seed, downloads the corresponding Web page, extracts the hyperlinks contained in it, and recursively continues to download the Web pages referred by the hyperlinks. After extracting a given hyperlink from a page, the crawler checks whether the URL has been already visited, or not, and only nonvisited URLs are added to the list of URLs to be crawled. The contents of a Web page are then sent to the OpenCalais and to the AlchemyAPI Web services for its semantic annotation.

OpenCalais processes a given Web page and returns annotated semantic metadata containing identified entities, facts, events, topics and social tags as RDF payloads serialized as XML data. Entities are classified into people, companies, industry terms, political event, position, location, technology, and product. Events and facts such as acquisitions, management changes in companies, alliance, bankruptcy, credit rating, and joint venture are recorded in the metadata. This metadata also contains all relations that involve at least one recognized entity from the content. Relations are generally all subject-predicate-object relationships without predefining their types. In addition, the metadata includes the topics that content discusses. Finally, the metadata induced by OpenCalais also includes a feature called social tags which is an emulation of how a human being would tag specific pieces of content.

Web pages are also processed by the Alchemy API to generate complementary semantic metadata. The AlchemyAPI Web service processes a Web page text and

[a] http://www.opencalais.com.
[b] http://www.alchemyapi.com.

returns keywords, search terms and concept tags. It employs sophisticated statistical algorithms and NLP technology for the extraction of keywords and search terms from the content. The concept tagging feature of AlchemyAPI automatically tags documents and text in a manner similar to human-based tagging. The results are returned as RDF payloads serialized as XML data.

The resulting XML data is parsed by SVLMC to extract the metadata and store it in the RDF data store. This semantic metadata is used to determine whether two Web pages are considered to be sufficiently semantically similar to make a RDF statement about their similarity. To measure semantic similarity between Web pages, we have used two methods based on the number of identical semantic metadata items between the Web pages. In the first method, similarity is calculated by simply counting the number of common semantic metadata items and the weight given to different semantic metadata items is equal. The second method makes use of a similarity coefficient that gives different weights to semantic components according to their frequency. The reasons that justify these approaches are the ease of implementation and the accurate computation of semantic similarity score values. These methods will be able to generate accurate results because the more similar semantic components two pages have the more likely it is that they are similar. For example, two pages of the same topic may be similar, but they are even more similar if they are about the same topic and have the same keywords. In the following subsections, we will discuss the two similarity computation methods.

3.3.1. *Similarity score using equal weights*

The similarity between any two Web pages is computed by comparing their semantic information. In this method, equal weight is given to each semantic metadata component. A similarity score is computed that is incremented for every two semantic items that are identical. Let S_1 and S_2 be set of semantic metadata items in the Web pages P_1 and P_2, respectively. The similarity score is given by,

$$\text{Similarity Score}(P_1, P_2) = |S_1 \cap S_2|. \tag{4}$$

The semantic similarity matrix M is generated in such a way that each entry of M is the normalized similarity score value between the two Web pages. The semantic similarity matrix M is normalized in such a way that each entry is between 0 and 1. To attain this normalization, each entry in the row is divided by row sum.

We now give an example to illustrate the score computation. Consider a case in which two Web pages are compared for semantic similarity. The RDF output for the first URL contains one topic (Technology Internet), one social tag (Computer Network), keyword (Semantic Web) and search term (Technology). In addition, the RDF output for the second URL contains the topic (Technology Internet), the social tag (Computer Network), the keyword (Semantic Web) and the search term (Java).

The system compares topics from both the URLs and finds one match, hence the score is incremented by one. Then the system compares social tags and finds one match, hence the score is incremented by one. Also system compares keywords and search terms and finds one match. The similarity score in this case is three. If we assume that preset threshold values for low and high similarity are 2 and 10, respectively, the similarity will be medium similarity, since similarity score value is between two thresholds.

3.3.2. *Similarity using different weights*

This method is an extension of method discussed in Sec. 3.3.1. A Web page is represented by an n-dimensional vector v, where n is the total number of semantic metadata items in the Web site. The vector for each Web page 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 weight of the jth semantic metadata item f_j in the page p. The term frequency and inverse document frequency (TF-IDF) is used to produce a composite weight for each semantic metadata in a Web page. The term frequency of semantic metadata item f_j is given by[38]

$$\mathrm{TF}_{f_j,p} = \frac{\text{frequency of semantic metadata item } f_j \text{ in Web page } p}{N_p}, \tag{5}$$

where N_p is the total number of semantic metadata items in the Web page p. The inverse document frequency of semantic metadata item f_j is given by[38]

$$\mathrm{IDF}_{f_j} = \log \frac{N}{df_j}, \tag{6}$$

where N is the total number of Web pages in the Web site and df_j is the document frequency of the semantic metadata item f_j i.e., the number of Web pages in the Web site containing the semantic metadata item f_j. The weight $w(f_j, p)$ of the semantic metadata item f_j in page p is given by the TF-IDF weighting scheme,[38]

$$w(f_j, p) = \mathrm{TF}_{f_j,p} \times \mathrm{IDF}_{f_j}. \tag{7}$$

Let P_1 and P_2 be two Web pages being compared. We make use of the Tanimoto coefficient[39] to calculate the similarity between the two Web pages, which is an extended version of the Jaccard Coefficient and cosine similarity. It assumes that each data object is a vector of attributes and that such attributes may, or may not be binary. In case all the attributes are binary, the Tanimoto method corresponds to the Jaccard method. The Tanimoto coefficient has been widely used to calculate similarity between documents in text mining and is computed as defined from the following equation,

$$\text{Similarity Score}(P_1, P_2) = \frac{\sum_{j=1}^{n} w(f_j, P_1) w(f_j, P_2)}{\sum_{j=1}^{n} w(f_j, P_1)^2 + \sum_{j=1}^{n} w(f_j, P_2)^2 - \sum_{j=1}^{n} w(f_j, P_1) w(f_j, P_2)}, \tag{8}$$

where $w(f_j, P_1)$ and $w(f_j, P_2)$ are the weights of the semantic metadata component f_j in the two pages P_1 and P_2, respectively. The similarity score is the dot product of the

Table 3. Similarity computation using different weights.

Vector	Semantic metadata items				
	Internet technology	Computer networks	Semantic web	Technology	Java
TF($P1$)	0.20	0.20	0.20	0.20	0.00
TF($P2$)	0.20	0.20	0.20	0.00	0.20
IDF	0.54	0.54	0.54	0.84	0.84
V_{p1}	0.11	0.11	0.11	0.17	0.00
V_{p2}	0.11	0.11	0.11	0.00	0.17

vectors divided by the squared magnitudes of the vector minus the dot product. The semantic similarity score will have a value between 0 and 1, where 1 means that the Web pages have exactly the same semantic metadata items and 0 means that there are no similar semantic metadata items. The semantic similarity matrix M is generated using the similarity score value in such a way that each entry of M is the similarity score value between the two corresponding Web pages.

Table 3 shows the similarity computation using the TF-IDF scheme discussed above. In the previous example the total number of semantic metadata items is five. In this example, for the sake of illustration, we have assumed that the total number of Web pages in the Web site is seven. The value of the total number of pages is required for the computation of the IDF vector as shown in Eq. (6). One additional assumption is that the semantic metadata items appearing in these two Web pages do not appear in the other remaining Web pages. Let TF($P1$) and TF($P2$) represents term frequency vectors for two pages $P1$ and $P2$ being compared. The two Web pages are represented by the vectors V_{p1} and V_{p2}. By using Eq. (8) and vectors V_{p1} and V_{p2} the value of the similarity score between pages $P1$ and $P2$ will be 0.38. If we assume that the values for the low and high similarity preset threshold are 0.22 and 0.37, respectively, the similarity score will be high, since the value is above the threshold for high.

The semantic similarity score is used to determine whether two Web pages are considered to be sufficiently semantically similar to make a RDF statement about their similarity. The similarity will be considered low, medium or high depending on its comparison to preset thresholds. If the score is below the low threshold, then the semantic similarity between the pages will be considered low, and if the score is between the low and high threshold then similarity will be medium. The score will be considered high if it is above the high threshold. These preset threshold values are dependent on the data set in question. The thresholds are decided based on average similarity score value between all pages in a Web site under consideration. Let $P = \{p_1, p_2, \ldots, p_n)$ be set of all Web pages in the site under consideration, the average similarity score value between all pages is calculated by using following equation,

$$S_{\text{avg}} = \frac{\sum_{p_x, p_y \epsilon P} \text{Similarity}(p_x, p_y)}{|P|}. \tag{9}$$

Based on average similarity score value, preset thresholds for low and high similarity are calculated by using following equations,

$$T_{\text{low}} = 0.75 * S_{\text{avg}}, \tag{10}$$

$$T_{\text{high}} = 1.25 * S_{\text{avg}}. \tag{11}$$

By using thresholds for low and high similarity, the similarity between Web pages is determined by using the following equation,

$$\text{Similarity}(P_x, P_y) = \begin{cases} \text{low} & \text{if similary score} \leq T_{\text{low}}, \\ \text{medium} & \text{if similary score} > T_{\text{low}} \text{ and } \leq T_{\text{high}}, \\ \text{high} & \text{if similary score} > T_{\text{high}}. \end{cases} \tag{12}$$

The SVLMC system is automated to compare every pair of Web pages in a Web site being analyzed. The semantic similarity score is used to combine semantic information with usage data in VLMC model as discussed in Sec. 3.5. In addition, the semantic similarity information such as high, medium, or low between the pages is used for generating semantic recommendations discussed in Sec. 3.6.

3.4. *RDF store*

A RDF Data Store, also called a Semantic Repository, is able to store huge volumes of RDF data, perform necessary inference according to semantics of the data and provide a powerful query-answering mechanism that operates in real time. A key benefit of RDF Data Store is the ability to perform SPARQL queries, eliminating the need for custom coding. The SVLMC makes use of AllegroGraph RDF Data Store.[c] The AllegroGraph RDF Store is a modern, high-performance, persistent RDF graph database.

3.5. *Semantic rich variable length Markov model*

The analysis of the contents of user visited Web pages is a natural step to better capture users' information goals. Thus we enhance, VLMC model in order to take into account Web page contents. There are several alternative ways to integrate semantic knowledge into Markov Chain models: one approach is to integrate it in the Transition Probability Matrix as discussed in the following subsections. The reason behind choosing this approach is it is very easy to integrate usage and semantic data because, link probability in the VLMC model and semantic similarity have been represented as a matrix.

A transition probability in the VLMC model is represented by a transition probability matrix T in which an entry $T_{i,j}$ represents the transition probability between pages i and j. For the integration of semantics into a VLMC model, we have used the approach presented in Ref. 6. In that approach, the semantic information used to characterize Web pages is obtained from the domain ontology provided by

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

the ontology engineer during the design of a Web site. The authors have assumed that a single Web page represents only one concept from the ontology, which is not always the case in real world. The semantic distance between two pages is calculated based on number of edges separating two pages in the domain ontology.

The SVLMC model makes use of semantic metadata to calculate the semantic similarity instead of distance in domain ontology used in Ref. 6. Since a Web page may represent more than one concept, we use the methods described in Sec. 3.3 to represent the similarity between pages. 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 M is combined with the transition matrix T of a VLMC model in order to derive the weight matrix W by using,

$$W_{p_i,p_j} = \alpha * T_{p_i,p_j} + \begin{cases} (1-\alpha) * M_{p_i,p_j}, & M_{p_i,p_j} > 0, \\ 0, & M_{p_i,p_j} = 0, \end{cases} \tag{13}$$

where α is the semantic combination factor (SCF) that balances the two measures, link probability in the VLMC model and semantic similarity score between the Web pages. By varying SCF, it is possible to tune performance of the system by giving more weight to the usage data or to the semantic information. We note that, in the conducted experiments (see Sec. 4.2), the recommendation accuracy is sensitive to α. It is observed that recommendation accuracy does not vary significantly for the value of α in the range of 0.3 to 0.7 and that a value of $\alpha = 0.5$ is adequate in general.

In order to illustrate the concept, Table 4 shows the transition probability matrix corresponding to the first-order model given in Fig. 2. Table 5 depicts the semantic similarity score matrix M calculated using similarity computation method discussed in Sec. 3.3.2, and Table 6 gives the resulting weight matrix W. All these tables include fictitious start and end states added during construction of the first-order model. The state 0 stands for the start state (S) and state 8 for the last state (F). For this example, the value of SCF α is set at 0.5. For computation of the semantic similarity between page 1 and page 3, we note that the transition probability and semantic similarity between the two pages are 0.56 and 0.35, respectively. Given the

Table 4. Transition probability matrix T.

	0	1	2	3	4	5	6	7	8
0	0.00	0.56	0.44	0.00	000	0.00	0.00	0.00	0.00
1	0.00	0.00	0.00	0.56	0.38	0.06	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.80	0.00	0.00	0.20	0.00
3	0.00	0.00	0.00	0.00	0.00	0.58	0.42	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.30	0.46	0.24
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 5. Semantic similarity matrix M.

	0	1	2	3	4	5	6	7	8
0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.00	0.00	0.38	0.35	0.19	0.31	0.22	0.22	0.00
2	0.00	0.38	0.00	0.40	0.19	0.38	0.20	0.20	0.00
3	0.00	0.35	0.40	0.00	0.16	0.35	0.20	0.10	0.00
4	0.00	0.19	0.19	0.16	0.00	0.32	0.24	0.30	0.00
5	0.00	0.31	0.38	0.35	0.32	0.00	0.32	0.20	0.00
6	0.00	0.22	0.20	0.20	0.24	0.32	0.00	0.55	0.00
7	0.00	0.22	0.20	0.10	0.30	0.20	0.55	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 6. Combined matrix W.

	0	1	2	3	4	5	6	7	8
0	0.00	0.28	0.22	0.00	0.00	0.00	0.00	0.00	0.00
1	0.00	0.00	0.19	0.46	0.29	0.19	0.11	0.11	0.00
2	0.00	0.19	0.00	0.20	0.50	0.19	0.10	0.20	0.00
3	0.00	0.18	0.20	0.00	0.08	0.47	0.31	0.05	0.00
4	0.00	0.10	0.10	0.08	0.00	0.16	0.27	0.38	0.12
5	0.00	0.16	0.19	0.18	0.16	0.00	0.16	0.10	0.50
6	0.00	0.11	0.10	0.10	0.12	0.16	0.00	0.28	0.50
7	0.00	0.11	0.10	0.05	0.15	0.10	0.28	0.00	0.50
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

SCF of 0.5, the entry for the combined matrix W will be $0.5 * 0.56 + (1.0 - 0.5) * 0.35 = 0.46$. In the proposed SVLMC, the weight matrix W is used for the next link of choice prediction and for the recommendation generation, instead of transition probability matrix T used by the standard VLMC model.

3.6. *Recommendation engine*

Web recommendation is a promising technology that attempts to predict interests of Web users by providing users information and/or services that they need without the users explicitly asking for them. In the proposed SVLMC, recommendation engine makes use of model discussed in Sec. 3.5 and of the semantic metadata stored in RDF Data Store to generate recommendations. The use of semantic metadata can provide significant advantages in recommendation and personalization.

3.6.1. *Recommendations using semantically enriched transition probability matrix*

In the proposed SVLMC model, the transition probably matrix of the VLMC model is enriched with semantics present in Web page contents, as described in Sec. 3.5. The semantically enriched matrix W is used to predict the next link choice in order to generate recommendations.

The prefix path followed by a Web user is matched into the semantically enriched matrix W in order to generate the next pages' recommendations. This technique is called popular path prediction.[26] In this process, the most probable page from the tip of the path is given as a prediction and added into the prefix path. This process is repeated until the required size recommendation set is generated or further pages cannot be predicted by the SVLMC model. If the number of recommended pages is less than the size of recommendation set, then predicted pages are semantically expanded in order to generate required size recommendation set.

3.6.2. *Semantic recommendations*

In order to generate semantic recommendations, the system takes into account the past navigation history modeled in terms of VLMC Model and the Web page contents modeled in terms of RDF Data.

In this case, for prediction of the next link of choice, the current user's session is completely or partially matched with the VLMC model and the two most probable pages are provided as predictions. These predicted pages are then semantically expanded and recommendations are generated. The semantic expansion consists of identifying a set of pages that are semantically similar to the predicted pages, pages in a Web site that discuss the topics contained in the predicted pages, and the pages having similar social tags or concepts to those in the predicted pages.

In case the user's session is not found in the VLMC model, that is, that particular path was not traversed before according to the history data, then, based on the Web pages in the session, the user's topics of interest are identified and the RDF Data Store is queried to retrieve related pages. In this way, SVLMC model addresses the problem of unseen data of the VLMC model.

To illustrate the recommendation methods, consider a navigation session $\langle 1,3,5,6\rangle$. For illustration purposes, assume that the recommendation set size is two and the surrogate active session derived from this session is $\langle 1,3\rangle$. This surrogate active session is fed to the first-order VLMC model discussed in Sec. 3.2.1 and the recommendation methods proposed in this section are applied. The prefix path $\langle 1,3\rangle$ is matched in the VLMC model and the generated recommendations set is $\langle 5\rangle$. Further recommendations cannot be generated from state 5. The recommendations generated using the method discussed in Sec. 3.6.1 is $\langle 2,5\rangle$. In case of recommendation method discussed in Sec. 3.6.2, the prefix path $\langle 1,3\rangle$ is matched in VLMC model and the predicted pages 5, 6 are semantically expanded to generate recommendation set $\langle 5,6\rangle$. In this case, semantic expansion is not needed since the recommendation set size is two, which is the same as the number of predicted pages.

4. Experimental Evaluation

In this section, we describe the experiments conducted and discuss their results. In Sec. 4.1, we state the purposes of the experiments, then the experimental design, the data sets and the evaluation metrics. In Sec. 4.2, we present the experimental results and its discussion.

4.1. *Experiments setup*

4.1.1. *Purposes of experiments*

The SVLMC model combines usage data and detailed Web site semantic contents. It is often observable that the Web pages a visitor requests during his navigation within a Web site are correlated and can be bound to a specific content topic. We therefore hypothesize that by incorporating semantic Web page contents into the model it is possible to capture users' navigation goals more precisely and, therefore, increase prediction and recommendation accuracy of the system. In order to assess the validity of our hypothesis, the SVLMC model will be evaluated in terms of precision, coverage, F1 measure and R measure.[40] In addition, in order to assess the advantages of using detailed semantic data, the SVLMC model is compared to the solely usage-based VLMC model and to Markov models enriched with semantic data expressed solely by means of ontology terms, as is the models presented in Ref. 6.

In order to implement the first model presented in Ref. 6, Markov models from first to fourth order were built for each data set. Web pages in the site were mapped to the classes in the ontology and the semantic distance was calculated based on number of edges between the classes of corresponding Web pages in the ontology. Thus, semantic information in the form of a semantic distance matrix was computed. The semantic distance matrix was directly combined with the transition probability matrix of a Markov model and the resulting matrix is used to generate the recommendations.

The second semantically enriched Markov model presented in Ref. 6 makes use of semantic distance as a measure to prune the states in a SMM. First an All-Kth-Order Markov model proposed in Ref. 41 is built and then the states in the model which have zero frequency are pruned. If the semantic distance between the pages is higher than the maximum allowed semantic distance, the states are pruned to reduce the state space complexity of the All-Kth-order model. The resulting semantic-pruned model is used to generate the recommendations. Further details on the pruning process are available in Ref. 6.

Several experiments were conducted to assess the performance of the proposed methods. The questions to be answered through the experimentation are as follows:

(1) Does integration of content results in improvement in recommendation accuracy?
(2) Does semantically enriched lower order VLMC model provides better prediction and recommendation accuracy compared to higher order VLMC model and to models that makes use of semantic data expressed by means of ontology terms?
(3) Which of the two proposed semantic integration methods is able to provide more accurate recommendations?

(4) What is the value of SCF α which leads to accurate recommendations? And does the value of α depends on the data set used?

4.1.2. *Experiment design*

To address the stated questions, a set of experiments were performed for different values of the parameters of the system. 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. It was observed experimentally that $k = 5$, is adequate for the data sets considered. In order to simulate active sessions of a Web user, concept of sliding window over the sessions was used. We let w denote the number of pages taken from a test session being processed. For a given session, and for a given w, a sequence of w Web pages is selected as a surrogate active session. Each active session was then fed into the model in order to produce a recommendation set. The recommendation set obtained is then compared to the remaining items in the test session in order to compute precision, coverage, $F1$ measure, and R scores.[40] This recommendation set was varied from 6 to 10 Web pages for each surrogate active user session. For each of these measures, final score for a particular test session was the mean score over all of the surrogate active sessions associated with the session. Finally, the mean over all sessions in the testing set was computed as the overall evaluation score for each measure. All experiments were conducted on a Pentium IV Dell Desktop machine. The model has been implemented in Java, and makes use of MS SQL Server 2008 database to store application data.

4.1.3. *Data sets*

We note that in order to conduct the SVLMC experimental evaluation, it is necessary to use a data set providing both Web server log data and Web page contents. These experiments have been conducted on the publicly available Music Machine data set (DS-1), on the Semantic Web dog food Web site (DS-2), and on a synthetic usage data generated for the technical university Web site (DS-3).

For the DS-1 data set, we have used access entries in the three month period, February to April 1999. This data was made available with requests organized in sessions. The data set is available at http://www.cs.washington.edu/ai/adaptive-data/.

For DS-2 we have used access entries between August 2010 and December 2010 from the Semantic Web Dog Food Web site. This is a constantly growing Web site of publications, people and organizations in Web and Semantic Web area, covering several of the major conferences and workshops in the field. In case of DS-2, in order to avoid a large sparse transition probability matrix, a filtering step was applied in order to remove the least requested pages. This data set is available at http://data.semanticweb.org/usewod/2011/dataset2/usewod2011_dataset.tar.

The DS-3 makes use of a university Web site including Web pages of individuals (i.e., students and teachers), news group and courses correspond to average values for which usage data was generated using the technique described in Ref. 42.

Table 7. Statistics of experimental data set.

Attributes	DS-1	DS-2	DS-3
Total access entries	539,824	252,192	405,147
Total clean access entries	539,824	230,252	405,147
Total accessed Web pages in log	1,540	1,919	950
Total pages identified by crawler	1,715	2,105	1,200
Different access users	27,500	6,245	25,000
Total identified sessions	29,220	7,667	25,000
Total identified sessions ($\geq$2 requests)	27,934	7,032	25,000
Number of RDF triples	1,387,661	4,983,594	642,573

Table 7 depicts statistics of the experimental data sets. For each data set, total access entries, clean access entries after removing entries that are not useful to represent user Web navigation behavior, number of pages occurring in the log, total pages identified by crawler during crawling of a Web site and total users identified are indicated. Also total number of sessions derived from each data set and the number of sessions of lengths more than two; session length is measured by number of requests a session is composed of, are given. It is assumed that identified sessions having session length more than two pages are more suitable for our experiments, since it might carry more information about Web users' intention on a Web site. Therefore, sessions having more than two page requests are chosen for the experimental evaluation. Table 7 also indicates number of RDF triples identified from each Web site.

As described in Sec. 4.1.1, the SVLMC methods are compared with two semantic ontology methods proposed in Ref. 6 that make use of the ontology of a Web site. Ontology is a formal, explicit specification of a conceptualization,[43] which can capture reusable knowledge in a domain. As an explicitly defined and machine-processable abstract model, ontologies were developed for the purpose of knowledge sharing to provide common understanding about domain knowledge. Ontology building is a tedious process that requires much time and resources. For the three data sets used in the experiments, we constructed the ontology semi-automatically by using Text2Onto[44] and the Protégé[d] tool. The concept represented by a Web page is identified by using Text2Onto. The concepts are represented as classes and subclasses in the ontology. Relations, also called properties, are used to depict the hierarchy of the classes in the ontology. The user interface provided by Protégé is used to construct the ontology. The relations used in constructing ontology are like isPartOf, isSubclassOf, subTopicOf, isOrganizedBy, presentedBy, taughtBy, isManufacturedBy and facilatedWith. For example, in case of the ontology for DS-1, the relation used is isManufacturedBy to relate the concept manufacturer and its instance.

[d] http://protege.stanford.edu.

Table 8. Complexity of ontology.

Description	DS-1	DS-2	DS-3
Total concepts	253	200	408
Total root classes	75	20	28
Total leaf classes	197	161	339
Total relations	8	10	12
Average depth of inheritance tree	1.38	2.83	4.1
Edge node ratio (ENR)	0.72	1.46	2.57

Table 8 shows the complexity of the ontology generated for each of the three data sets. For each data set, the total number of concepts, root classes,[45] leaf classes,[45] relations, the average depth of inheritance,[45] and the edge node ratio[46] are indicated.

Both methods proposed in Ref. 6 make use of the distance in the ontology as a measure of semantic similarity between Web pages. Web pages are mapped to their corresponding classes in the ontology and the level of the concept in the ontology hierarchy is used to calculate distance in the ontology.

4.1.4. *Evaluation metrics*

In order to evaluate effectiveness of the SVLMC recommendations, four standard measures were used: precision, coverage, $F1$ measure, and R measure.[40] Among these, precision and coverage metrics have been widely used in recommender system research. While, precision measures the degree to which a recommendation engine produces accurate recommendations, coverage measures the ability of a recommendation engine to recommend all pages that are likely to be visited by a user. As precision and coverage are inversely related, a combination measure named $F1$ is used that gives equal weight to precision and coverage.

Follows the definition of the metrics used. Assume that we have user session t from the test set viewed as a set of Web pages, and that we use a window $w \subseteq t$ of size $|w|$ to produce a recommendation set Rec using recommendation methods discussed in Sec. 3.6. Then precision of Rec with respect to t is defined as,[40]

$$\text{precision}(\text{Rec}, t) = \frac{|\text{Rec} \cap (t - w)|}{|\text{Rec}|}, \tag{14}$$

where $|\text{Rec} \cap (t - w)|$ is a number of common Web pages in both recommendation set and evaluation set. Precision is a number of relevant Web pages retrieved divided by the total number of Web pages in recommendations set.

Coverage of Rec with respect to t is defined as,[40]

$$\text{coverage}(\text{Rec}, t) = \frac{|\text{Rec} \cap (t - w)|}{|t - w|}. \tag{15}$$

Coverage is the ratio between the number of relevant Web pages retrieved and the total number of Web pages that actually belongs to the test user session. To evaluate the performance of the recommendation system, neither of these measures individually are sufficient to evaluate the performance. Ideally, one would like to obtain high precision and high coverage. A single measure that captures this is the $F1$ measure defined as,[40]

$$F1(\text{Rec}, t) = \frac{2 \times \text{precision}(\text{Rec}, t) \times \text{coverage}(\text{Rec}, t)}{\text{precision}(\text{Rec}, t) + \text{coverage}(\text{Rec}, t)}. \tag{16}$$

The $F1$ measure gains its maximum value when both precision and coverage are maximized. An additional measure, called the R measure, is obtained by dividing the coverage by the size of the recommendation set defined as,[40]

$$R(\text{Rec}, t) = \frac{\text{coverage}(\text{Rec}, t)}{|\text{Rec}|}. \tag{17}$$

The above four metrics, precision, coverage, $F1$ measure and R measure were used to assess the SVLMC performance.

4.2. *Experimental results*

Experiments were conducted in order to evaluate performance of the SVLMC model as compared to a solely usage-based VLMC model[7] and models that make use of semantic data expressed by means of ontology terms.[6]

In Sec. 3.3, two methods for computing the similarity between Web pages are proposed. The first method gives equal weights to different semantic metadata items and is based on simple counting the number of common semantic metadata items. The second method makes use of a similarity coefficient and different weight is given to semantic components. We will analyze impact of these similarity methods on the recommendation accuracy of SVLMC for the two the recommendation methods proposed in Sec. 3.6. Figure 4 gives the performance of the SVLMC, measured by $F1$ measure, for the two similarity methods for the three data sets DS-1, DS-2 and DS-3. In the following, SVLMC_tm_sw is the recommendation method proposed in

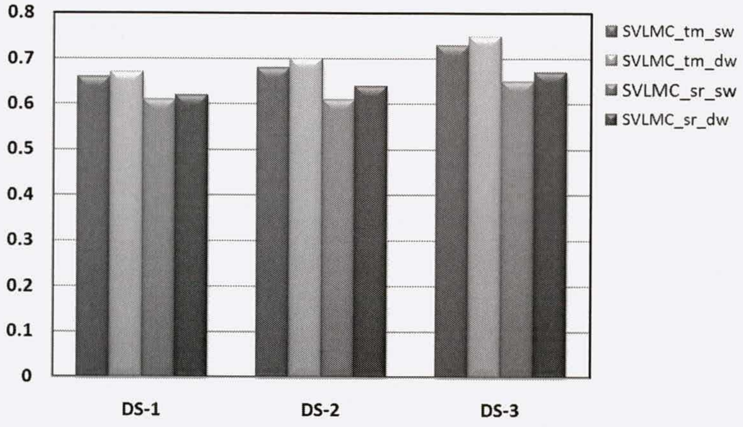

Fig. 4. Comparison of similarity computation methods.

Sec. 3.6.1 (transition matrix) while making use of the same weight (SW) similarity method proposed in Sec. 3.3.1 and SVLMC_tm_dw while making use of the different weights (DW) similarity method proposed in Sec. 3.3.2. The method SVLMC_sr_sw is the method proposed in Sec. 3.6.2 (semantic recommendation) with SW and SVLMC_sr_dw with DW. A third-order SVLMC is used for the analysis. It is observed that the DW similarity method achieves 1–2% performance improvement over the SW for both the recommendation methods. Hence for further experimentation of the SVLMC, we will use the similarity method that applies different weights to semantic metadata components.

We will now assess the impact of the SCF α on the recommendation method proposed in Sec. 3.6.1 which uses the combined matrix W. In order to determine the recommendation accuracy, we let the SCF α, which balances the weight given to the usage data and to the semantic information, vary from 0.0 to 1.0. Figures 5–7 show the performance of the SVLMC, measured by the $F1$ measure, for the different values of α for the three data sets DS-1, DS-2 and DS-3. The results indicate that the performance of the SVLMC model depends on the value of α. In fact, the performance decreases for values of $\alpha < 0.3$ (gives more weight to semantic similarity) and $\alpha > 0.8$ (gives more weight to transition probability). Similar results were obtained for the other two data sets. The performance of the SVLMC does not vary significantly for values of α between 0.3 and 0.7, indicating that any value of α in that range is adequate. Hence, the value $\alpha = 0.5$ was adopted for the other experiments. It is imperative, for a good performance of SVLMC, to have a balanced combination of

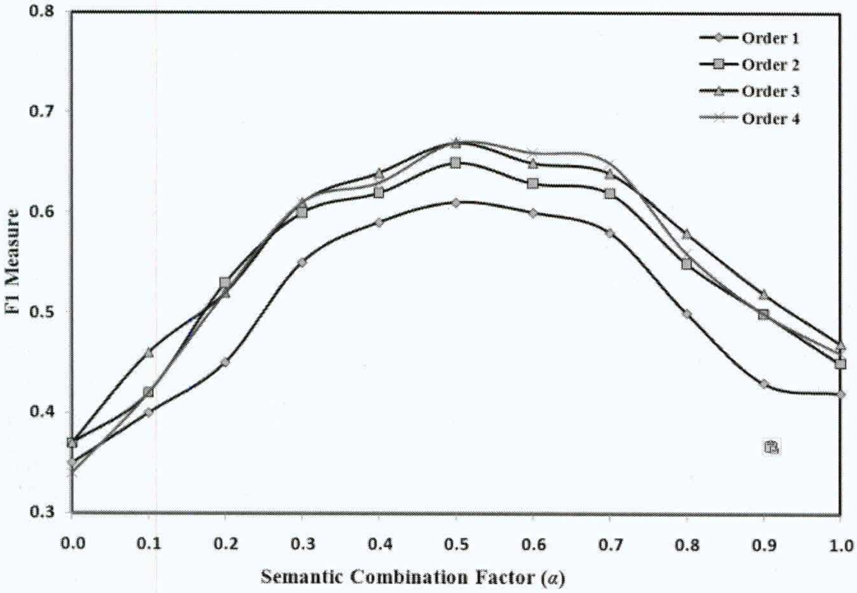

Fig. 5. Comparison of recommendation accuracy for different values of α on Music Machine Data Set (DS-1).

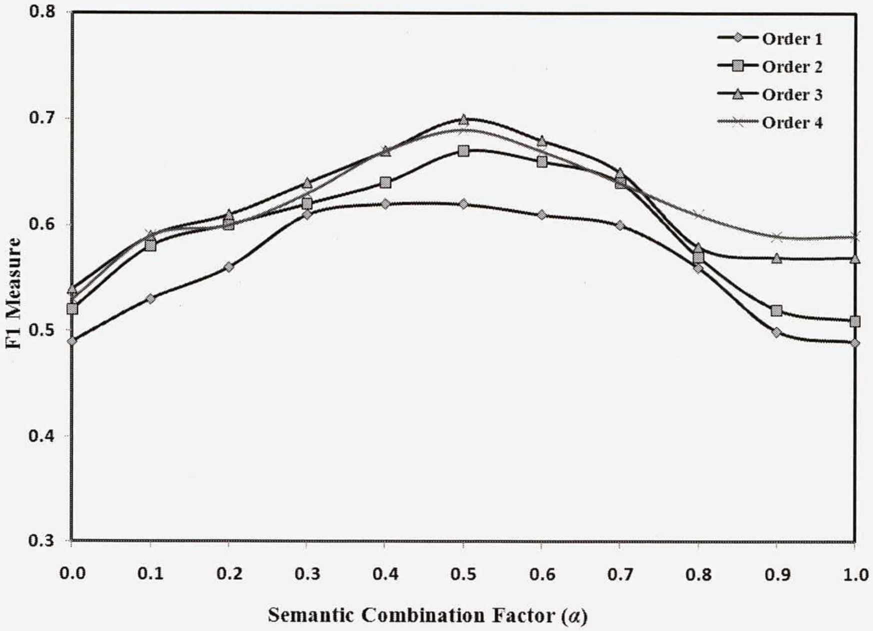

Fig. 6. Comparison of recommendation accuracy for different values of α on Semantic Web dog food Web site Data Set (DS-2).

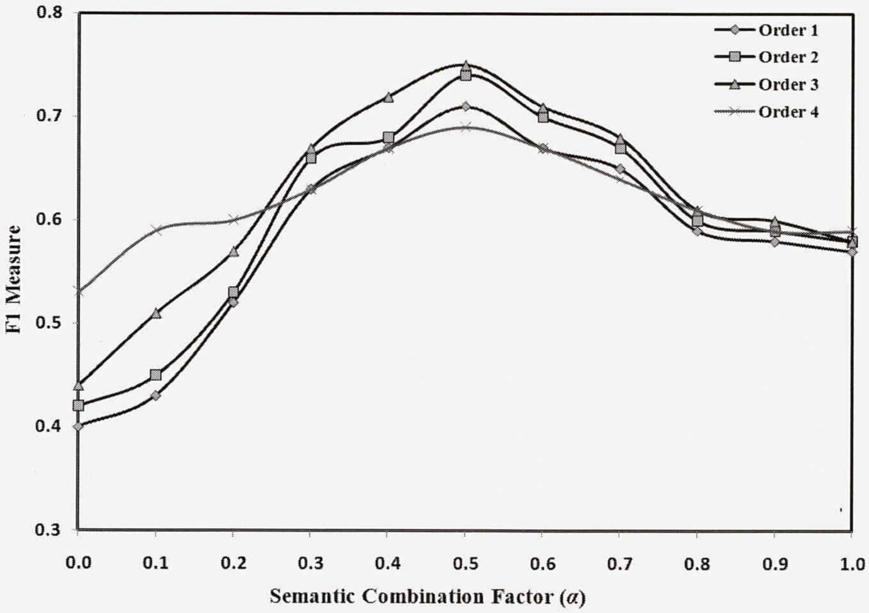

Fig. 7. Comparison of recommendation accuracy for different values of α on synthetic usage Data Set (DS-3).

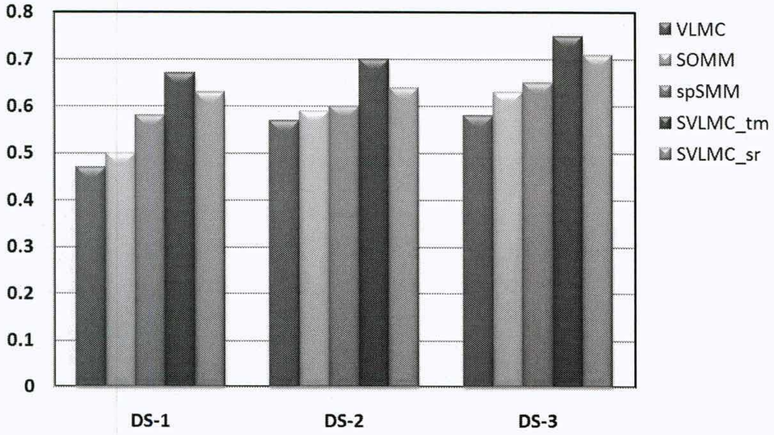

Fig. 8. Comparison of recommendation accuracy of proposed methods with VLMC and ontology-based models.

usage data and semantic metadata. Also the range of values of α for which SVLMC gives better performance is same for the three data sets and does not depend on the data sets used.

Figure 8 shows a comparison of the recommendation accuracy of VLMC model, the semantic ontology Markov model[6] (SOMM), semantic-pruned SMM[6] (spSMM), the more detailed semantic SVLMC model discussed in Sec. 3.6.1 (SVLMC_tm) and the semantic recommendation method discussed in Sec. 3.6.2 (SVLMC_sr). Again the metric used for the comparison is the $F1$ measure and a third-order Markov model is considered. The value of $\alpha = 0.5$ is adopted for the SVLMC model. It is observed from Fig. 8 that the SVLMC model outperforms the VLMC model and the SOMMs. Both SVLMC recommendation methods outperform the other models. The results reveal that the integration of more detailed semantic data characterizing the Web content has a positive impact on prediction accuracy of the model.

Table 9 shows more detailed comparison of results obtained for all the methods VLMC, SOMM, spSMM, SVLMC methods SVLMC_tm and SVLMC_sr, for the three data sets. For these experiments the value of α chosen is 0.5 for SVLMC_tm.

The results show that the SVLMC results are consistent for all the three data sets and that both the SVLMC recommendation methods outperform the VLMC model and both the SOMMs proposed in Ref. 6. Even for lower orders, SVLMC is able to achieve a performance comparable to a higher order VLMC model and to the semantic ontology models. Both the VLMC and the SVLMC models achieve the best accuracy for models of order 3. We will discuss the results of SVLMC_tm recommendation method. The SVLMC_tm achieves 15–20% performance improvement over the VLMC model, 8–15% performance improvement over SOMM, and 7–12% performance improvement over spSMM. The recommendation accuracy of the first-order SVLMC model is comparable to a third-order VLMC model. In case of DS-1, the number of states in the first-, second- and third-order models are 1715, 3126, and

Table 9. Results of recommendation engine for three data sets DS-1, DS-2, DS-3 using semantic integration method.

Method name	Model order															
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
	Precision				Coverage				*F*1 measure				*R* measure			
Data Set: Music Machine (DS-1)																
VLMC	0.41	0.44	0.49	0.46	0.49	**0.53**	**0.52**	**0.52**	0.42	0.45	0.47	0.46	0.04	0.04	0.05	0.05
SOMM	0.49	0.50	**0.57**	**0.56**	**0.51**	**0.54**	**0.54**	**0.53**	0.45	0.50	0.50	0.49	0.04	0.04	0.05	0.05
spSMM	**0.51**	**0.53**	**0.59**	**0.56**	**0.53**	**0.56**	**0.57**	**0.55**	**0.52**	**0.54**	**0.58**	**0.55**	0.04	0.04	0.05	0.05
SVLMC_tm	**0.60**	**0.65**	**0.71**	**0.70**	**0.60**	**0.63**	**0.62**	**0.64**	**0.63**	**0.65**	**0.64**	**0.65**	0.05	0.05	0.06	0.05
SVLMC_sr	**0.62**	**0.63**	**0.65**	**0.63**	**0.61**	**0.62**	**0.61**	**0.60**	**0.61**	**0.62**	**0.63**	**0.61**	0.05	0.05	0.06	0.05
Data Set: Semantic Web Dog Food Web site (DS-2)																
VLMC	0.40	**0.52**	**0.58**	**0.58**	**0.55**	**0.54**	**0.54**	**0.58**	0.49	**0.51**	**0.57**	**0.59**	0.04	0.05	0.05	0.05
SOMM	0.50	**0.54**	**0.60**	**0.60**	**0.55**	**0.55**	**0.55**	**0.58**	**0.50**	**0.52**	**0.59**	**0.59**	0.04	0.05	0.05	0.05
spSMM	**0.52**	**0.59**	**0.62**	**0.60**	**0.57**	**0.58**	**0.58**	**0.57**	**0.54**	**0.58**	**0.60**	**0.58**	0.04	0.05	0.05	0.05
SVLMC_tm	**0.62**	**0.69**	**0.73**	**0.72**	**0.60**	**0.64**	**0.65**	**0.64**	**0.62**	**0.66**	**0.67**	**0.66**	0.05	0.06	0.06	0.06
SVLMC_sr	**0.64**	**0.65**	**0.65**	**0.63**	**0.63**	**0.64**	**0.63**	**0.61**	**0.63**	**0.64**	**0.64**	**0.62**	0.05	0.06	0.06	0.06
Data Set: Synthetic (DS-3)																
VLMC	**0.65**	**0.65**	**0.64**	**0.64**	**0.53**	**0.54**	**0.54**	**0.54**	**0.57**	**0.58**	**0.58**	**0.58**	0.04	0.04	0.04	0.04
SOMM	**0.68**	**0.68**	**0.69**	**0.68**	**0.55**	**0.58**	**0.58**	**0.58**	**0.60**	**0.63**	**0.63**	**0.63**	0.04	0.05	0.05	0.05
spSMM	**0.69**	**0.69**	**0.70**	**0.67**	**0.59**	**0.58**	**0.60**	**0.59**	**0.63**	**0.63**	**0.65**	**0.63**	0.04	0.05	0.05	0.05
SVLMC_tm	**0.76**	**0.79**	**0.79**	**0.77**	**0.63**	**0.64**	**0.70**	**0.70**	**0.66**	**0.69**	**0.71**	**0.71**	0.05	0.05	0.06	0.06
SVLMC_sr	**0.70**	**0.70**	**0.71**	**0.69**	**0.62**	**0.64**	**0.65**	**0.62**	**0.65**	**0.66**	**0.67**	**0.65**	0.06	0.06	0.06	0.06

4507, respectively. Thus, a first-order SVLMC model is able to achieve same accuracy of a third-order VLMC model with a model size approximately 260% smaller, as measured by the number of states. For DS-2 data set, size of the first-, second- and third-order models are 2328, 3399 and 3736, respectively. The model size is reduced by 160%. Similarly for DS-3 data set, size of first-, second- and third-order models are 1200, 3025 and 5122, respectively. The model size is reduced by 420% if first-order model is used instead of third-order model. For all the three data sets, it was observed that average reduction in the model size is 280%. Thus SVLMC is able to manage tradeoff between size and accuracy of the model and lower order SVLMC models having high recommendation accuracy can be used.

It can be observed that the recommendations generated by SVLMC are better than those obtained with the VLMC model and SOMMs for all the three data sets. Thus, we conclude that the integration of semantic data induced from Web page contents is beneficial relatively to solely usage model and relatively to models using semantic ontology data.

The evaluation results for SVLMC_sr proposed in this paper are given in Table 9. It can be observed that the semantic recommendation method also outperforms the VLMC model and both the SOMMs. The accuracy of the SVLMC recommendation remains the same irrespective of the order of the model. The reason behind this is that the recommendations are generated based on semantic similarity rather than on transition probabilities (given by usage data) of the Web pages ahead in the path. In case of the semantic recommendation method, the first-order model seems to be sufficient to get good accuracy and there is no gain in using higher complexity models.

In the evaluation of a recommendation model, one difficulty emerges when a user session contains pages previously unseen by the model, that is, pages that are in the test set that did not occur in the training set. It was observed that the SVLMC model is able to provide recommendation set for previously unseen pages, since even though the transition probability is zero (given by the usage data), there will be semantic similarity score value present in the combined matrix W. Also it was observed that in case of many recommendation sets generated by SVLMC that contained many newly added pages, SVLMC model addressed new item problem of solely usage-based techniques.

In summary, analysis of the experimental results enables to answer the questions raised in Sec. 4.1.1. The detailed semantic integration of Web contents into recommendation process improves the recommendation results by 15–20% compared to solely usage-based VLMC model, 8–15% as compared to ontology-based model SOMM, and 7–12% as compared to spSMM.

- Integration of usage data and Web page content results in improved recommendation accuracy. In particular, detailed semantic data induced from the Web pages' content is better than semantic data based solely on the ontology concepts of a Web site.

- Semantically enriched lower order Markov models provide recommendation accuracy comparable to higher order models (a first-order SVLMC model is able to outperform a third-order VLMC).
- Semantic integration into the transition probability matrix outperforms over semantic recommendation method.
- The recommendation accuracy of SVLMC model depends on SCF α. Also it is observed that the value of α does not depend on the data set used. This is because for any data set, the proposed method gives better results when equal weightage is given to usage and semantic data.

5. Concluding Remarks

In this paper, a novel approach SVLMC is presented to predict users' future navigations requests. The model combines the VLMC[7,8] and detailed semantic information extracted from the Web pages. Two recommendation techniques were used. The first generates recommendations using a semantically enriched transition probability matrix, and the second method generates recommendations based on semantic data stored in RDF data store.

The proposed model is useful, for example, to generate a set of recommendations for a user at a very early stage of his interaction with the site, possibility leading to higher visitor retention and higher ratio for the conversion of casual browsers to potential customers in case of e-commerce sites. The SVLMC can be used to develop proper prefetching and caching strategies to reduce the Web server response time. The frequent navigation patterns generated by using SVLMC model can provide guidelines for improving navigation design of a Web site in order to satisfy Web user needs.

Extensive experiments were conducted in order to compare the proposed model to the usage-based Markov model (VLMC) and to Markov models enriched with semantic data from the ontology of concepts of the underlying Web site. The results show that user navigation behavior, derived from log files, combined with detailed semantic knowledge, can provide effective recommendations. The detailed semantic integration of Web contents into recommendation process improves the recommendation results by 15–20% compared to the solely usage-based VLMC model, 8–15% as compared to the ontology-based model, and 7–12% as compared to the spSMM. Thus, while previous works have shown the advantages of incorporating Markov models with semantic data based on the ontology of concepts of a Web site, herein we show that such models can be further enhanced by incorporating richer and more detailed semantic data induced from the actual Web pages' content. We further note that ontology-based models usually associate a single concept with each Web page while we enable a much richer characterization of each Web page.

In the last few years, size of dynamic Web pages is rapidly increasing. These dynamic pages cannot be indexed and processed easily. The proposed SVLMC method cannot be used for a database backed Web site that generates dynamic Web pages based on structured queries performed against backend databases. The contents of these Web pages depends on query parameters, hence these parameters must be taken into account in the personalization process. The SVLMC method can be extended to use with database backed Web site.

As future work, there are some aspects in which the proposed model can be improved. The model can be extended in order to take into account Web site structure. For example, a PageRank algorithm can be used to calculate the importance of the page within a Web site in order to enable recommendation engine to recommend Web pages having a degree of importance above a certain threshold. In addition, new methods to combine usage data modeled using VLMC model and semantic information can be devised. Finally, more sophisticated tools for modeling Web page contents with RDF can be applied to capture semantics in Web documents.

References

1. B. Liu, *Web Data Mining*, 2nd edn. (Springer, 2011).
2. V. N. Padmanabhan and J. C. Mogul, Using predictive prefetching to improve World Wide Web latency, *ACM SIGCOMM Computer Communication Review* **26**(3) (1996) 22–36.
3. J. Pei *et al.*, Mining sequential patterns by pattern-growth: The PrefixSpan approach, *IEEE Transactions on Knowledge Data Engineering* **16**(11) (2004) 424–1440.
4. F. M. Facca and P. L. Lanzi, Mining interesting knowledge from weblogs: A survey, *ACM Transactions on Data Knowledge Engineering* **53**(3) (2005) 225–241.
5. N. R. Mabroukeh and C. I. Ezeife, Using domain ontology for semantic Web usage mining and next page prediction, in *Proc. 18th ACM Conf. Information and Knowledge Management (CIKM 2009)* (2009), pp. 1677–1680.
6. N. R. Mabroukeh and C. I. Ezeife, Semantic-rich Markov models for Web prefetching, *IEEE Int. Conf. Data Mining Workshops* (2009), pp. 465–470.
7. J. Borges and M. Levene, Generating dynamic higher-order Markov models in Web usage mining, in *Proc. Ninth European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD)*, eds. A. Jorge, L. Torgo, P. Brazdil, R. Camacho and J. Gama (2005), pp. 34–45.
8. J. Borges and M. Levene, Evaluating variable length Markov chain models for analysis of user Web navigation sessions, *IEEE Transactions on Knowledge and Data Engineering* **19**(4) (2007) 441–452.
9. Y. Peng and K. Gang, A descriptive framework for the field of data mining and knowledge discovery, *International Journal of Information Technology & Decision Making* **7**(4) (2008) 639–682.
10. S. Oh and J. Kim, A sequence-element-based hierarchical clustering algorithm for categorical sequence data, *International Journal of Information Technology & Decision Making* **4**(1) (2005) 81–96.
11. X. Huang, Comparison of interestingness measures for Web usage mining: An empirical study, *International Journal of Information Technology & Decision Making* **6**(1) (2007) 15–41.
12. S. Park, N. Suresh and B. Jeong, Sequence-based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-means algorithm, *Data & Knowledge Engineering* **65** (2008) 512–543.

13. C. Makris, Y. Panagis, E. Theodoridis and A. Tsakalidis, A Web-page usage prediction scheme using weighted suffix trees, in *Proc. 14th Int. Conf. String Processing and Information Retrieval* (2007), pp. 242–253.
14. M. Jalali, N. Mustapha, M. N. Sulaiman and A. Mamat, WebPUM: A Web-based recommendation system to predict user future movements, *Expert Systems with Applications* **37** (2010) 6201–6212.
15. C. J. Carmona, S. Ramírez-Gallegoa, F. Torresb, E. Bernalc, M. J. del Jesusa and S. García, Web usage mining to improve the design of an e-commerce website: OrOliveSur.com, *Expert Systems with Applications* **39**(12) (2012) 11243–11249.
16. A. Guerbas, N. Salim, M. S. B. Ngadiman, W. Chimphlee and S. Srinoy, Effective Web log mining and online navigational pattern prediction, *Knowledge-Based Systems* **49** (2013) 50–62.
17. S. Jespersen, T. Pedersen and J. Thorhauge, Evaluating the Markov assumption for Web usage mining, in *Proc. Fifth ACM Int. Workshop Web Information and Data Management* (2003), pp. 82–89.
18. R. R. Sarukkai, Link prediction and path analysis using Markov chains, *Computer Networks* **33**(1–6) (2000) 337–386.
19. M. Levene and G. Loizou, Computing the entropy of user navigation in the Web, *International Journal of Information Technology & Decision Making* **2**(3) (2003) 459–476.
20. J. Borges and M. Levene, Comparison of scoring metrics for predicting the next navigation step with Markov model-based systems, *International Journal of Information Technology & Decision Making* **9**(4) (2010) 547–573.
21. W. K. Ching, E. S. Fung and M. K. Ng, Higher-order Markov chain models for categorical data sequences, *Naval Research Logistics* **51**(4) (2004) 557–574.
22. R. Leonardi, P. Migliorati and M. Prandini, Semantic indexing of soccer audio-visual sequences: A multi-modal approach based on controlled Markov chains, *IEEE Transactions on Circuits and Systems for Video Technology* **14**(5) (2004) 634–643.
23. F. Khalil, J. Li and H. Wang, A framework for combining Markov model with association rules for predicting Web page accesses, in *Proc. Fifth Australasian Data Mining Conf. (AusDM2006)*, Vol. 61 (2006), pp. 177–184.
24. S. Chimphlee, N. Salim, M. S. B. Ngadiman and W. Surat, Rough sets clustering and Markov model for Web access prediction, in *Proc. Post Graduate Annual Seminar* (2006), pp. 470–474.
25. M. Deshpande and G. Karypis, Selective Markov models for predicting Web page accesses, *ACM Transactions on Internet Technology* **4**(4) (2004) 163–184.
26. M. Eirinaki, M. Vazirgiannis and D. Kapogiannis, Web path recommendations based on page ranking and Markov models, in *Proc. Seventh ACM Int. Workshop Web Information and Data Management (WIDM '05)* (2005), pp. 2–9.
27. M. Eirinaki, D. Mavroeidis, G. Tsatsaronis and M. Vazirgiannis, Introducing semantics in Web personalization: The role of ontologies, in *Proc. EWMF/KDO'2005* (2005), pp. 147–162.
28. Z. Zhang and O. Nasraoui, Efficient hybrid Web recommendations based on Markov clickstream models and implicit search, in *Proc. IEEE/WIC/ACM Int. Conf. Web Intelligence*, Washington, DC, USA (2007), pp. 621–627.
29. H. Zhang, H. Song and X. Xu, Semantic session analysis for Web usage mining, *Wuhan University Journal of Natural Sciences* **12**(5) (2007) 773–776.
30. J. Hu and N. Zhong, Web farming with clickstream, *International Journal of Information Technology & Decision Making* **7**(2) (2008) 291–308.

31. A. C. M. Fong, B. Zhou, S. C. Hui, J. Tang and G. Y. Hong, Generation of personalized ontology based on consumer emotion and behavior analysis, *IEEE Transactions on Affective Computing* **3**(2) (2011) 152–164.
32. T. Nguyen, H. Y. Lu and J. Lu, Ontology-style Web usage model for semantic Web applications, in *Proc. 10th Int Conf. Intelligent Systems Design and Applications (ISDA)* (2010), pp. 784–789.
33. P. Senkul and S. Salin, Improving pattern quality in Web usage mining by using semantic information, *Knowledge and Information Systems* **30**(3) (2011), pp. 1–15.
34. M. Spiliopoulou, B. Mobasher, B. Berendt and M. Nakagawa, A framework for the evaluation of session reconstruction heuristics in Web-usage analysis, *INFORMS Journal on Computing* **15** (2003) 171–190.
35. R. Cooley, B. Mobasher and J. Srivastava, Data preparation for mining world wide Web browsing patterns, *Knowledge and Information System* **1** (1999) 5–32.
36. B. Berendt, B. Mobasher, M. Spiliopoulou and J. Wiltshire, Measuring the accuracy of sessionizers for Web usage analysis, in *Proc. Workshop on Web Mining at the First SIAM Int. Conf. Data Mining*, Chicago, IL, USA (2001), pp. 7–14.
37. J. Borges and M. Levene, Testing the predictive power of variable history Web usage, *Soft Computing* **11**(8) (2006) 717–727.
38. C. D. Manning, P. Raghavan and H. Schütze, *An Introduction to Information Retrieval* (Cambridge University Press Cambridge, England, 2008).
39. D. J. Rogers and T. T. Tanimoto, A computer program for classifying plants, *Science* **132** (1960) 1115–1118.
40. B. Mobasher, H. Dai, T. Luo and M. Nakagawa, Discovery and evaluation of aggregate usage profiles for Web personalization, *Data Mining and Knowledge Discovery* **6**(1) (2002) 61–82.
41. J. Pitkow and P. Pirolli, Mining longest repeating subsequences to predict www surfing, in *Proc. 2nd USENIX Symp. Internet Technologies and Systems*, Vol. 2 (1999), pp. 13–21.
42. P. I. Hofgesang and J. P. Patist, On modelling and synthetically generating Web usage data, *Int Conf. Web Intelligence and Intelligent Agent Technology* (2008), pp. 98–102.
43. T. R. Gruber, A translation approach to portable ontology specifications, *Knowledge Acquisition — Special issue: Current Issues in Knowledge Modeling* **5**(2) (1993) 199–220.
44. P. Cimiano and J. Völker, Text2Onto: A framework for ontology learning and data-driven change discovery, in *NLDB'05 Proc. 10th Int. Conf. Natural Language Processing and Information Systems* (2005), pp. 227–238.
45. H. Yao, A. M. Orme and L. Etzkorn, Cohesion metrics for ontology design and application, *Journal of Computer Science* **1**(1) (2005) 107–113.
46. H. Zhang, Y. Li and H. K. Tan, Measuring design complexity of semantic web ontologies, *The Journal of Systems and Software* **83**(5) (2010) 803–814.