A Personalized Product Based Recommendation System Using Web Usage Mining and Semantic Web

Sneha Y. S, G Mahadevan, and Madhura Prakash

***Abstract*—To be globally competent and competitive a successful presence on the web is necessary to sustain and retain itself in the market. The WWW is an interesting area for data mining because of abundance of information. Web users exhibit a variety of navigational interests through clicking a sequence of web pages. Analysis of this data will lead to discover many interesting patterns and facilitate users to locate more preferable web pages. Advanced mining processes are needed for this knowledge to be extracted, understood and used. Web Usage Mining (WUM) systems are specifically designed to carry out this task by analyzing the data representing usage data about a particular Web Site. The semantic information of the Web page contents is generally not included in Web usage mining. Online recommendation and prediction is one of the web usage mining applications. In this paper we present architecture for integrating semantic information about the products with web log data and generate a list of recommended products by using LCS Algorithm.**

***Index Terms*—WUM, LCS, Semantic Web, RDF, Recommendation.**

I. Introduction

Due to explosive growth of information over the internet in last several decades, information overload is becoming a big challenge. It has also become difficult for users to access relevant information efficiently. Meanwhile, the substantial increase in the number of websites presents a challenging task for webmasters to organize the contents of the websites to cater to the needs of users. Modeling and analyzing web navigation behavior is helpful in understanding what information online user's demand. Web Usage Mining (WUM) is the process of extracting knowledge from Web user's access data by exploiting Data Mining Technologies. It can be used for different purposes such as personalization, recommendation system improvement and site etc.

Since the web data is semi structured or structured a semantic knowledge of the data will be helpful in understanding the data. Semantic means that the meaning of data can be discovered by computers.

As defined by Tim Berners-Lee "The Semantic Web is an extension of the current web in which information is given a well-defined meaning, better enabling computers and people to work in cooperation. "The Semantic Web creates a universal medium for information exchange by putting documents with computer-process able meaning (semantic) on the World Wide Web. The Semantic Web extends the Web through the use of standards, markup languages and related processing tools.

Table 1. gives different WUM model and Semantic Based WUM Model.

TABLE I.

Sl. No.	WUM Model and Semantic Based WUM Model
1	Analog by Yan and Jacobsen
2	Web Personalizer by Mobasher, et al.
3	Liu and Kaselj
4	SUGGEST by Baraglia and Palmerini
5	OLRWUMS by Subhash Shinde, et al.
6	Recommendation System By Mehradad Jalali,et al.
7	Semantic based recommendation system by Suleyman et al
8	SWAPRS by Xin Sui, Suozhu Wang, Zhaowei Li

To overcome the drawbacks of the current recommender system such as intelligence, adaptability, flexibility, limitation of accuracy, we present architecture for integrating semantic information about the products with web log data and generate a list of recommended products by using LCS Algorithm.

The rest of this paper is organized as follows: In section 2, we review recent research advances in web usage mining and the comparative study of the different WUM systems and integrating of semantic information with web log data. Section 3 describes the main phases of the architecture and section 4 focuses on the LCS algorithm Section 5 focuses on personalizing the contents for the user and produces a list of recommended products. Finally, section 6 summarizes the paper and introduces future work.

II. Background and Related Works

Analog [1] is one of the first WUM systems. It is structured around an off-line and an online component. The off-line component builds session clusters by analyzing past users activity recorded in server log files. Then the online component builds active user sessions which are then classified according to the generated model. But this approach suffered from various limitations related to scalability and to the effectiveness of the results found.

Mobasher et al [2] presented Web Personalizer, a system which provides dynamic recommendations as a list of hypertext links to users. The analysis is based on anonymous usage data combined with the structure formed by the hyperlinks of the site. Data mining techniques are used in the preprocessing phase in order to obtain aggregate usage profiles. Web Personalizer is a good example of two-tier

Manuscript received February 16, 2012; revised March 27, 2012.

Sneha Y. S is with the Anna University of Technology, Coimbatore, India and Sr lecturer in Dept of CSE JSSATE, Bangalore, India (e-mail: sneha_girish@yahoo.com).

Dr. G. Mahadevan is with the AMCEC Bangalore, India (e-mail:g_mahadevan@yahoo.com).

architecture for Personalization systems.

Liu and Keselj [3] proposed the automatic classification of web user navigation patterns and proposed a novel approach to classifying user navigation patterns and predicting users' future requests. They used character N-grams to represent the contents of web pages, and combined them with user navigation patterns by building user navigation profiles composed of a collection of Ngrams.

Baraglia and Palmerini proposed a WUM system called SUGGEST that provide useful information to make the web user navigation easier and to optimize the web server performance [4]. SUGGEST adopts a two level architecture composed by an offline creation of historical knowledge and an online engine that understands user's behavior. Potential limitations of this architecture are: a) the memory required to store Web server pages is quadratic in the number of pages. This might be a severe limitation in large sites made up of millions of pages; b) it does not permit us to manage Web sites made up of pages dynamically generated.

The OLRWUMS is another WUM model proposed by Subhash Shinde et al., [5] short for the Online Recommendation for Predicting in Web Usage Mining system. They propose K-Nearest neighbor algorithm for classifying user's sessions.

Mehradad Jalali et al [6] proposed an Online Recommendation System using LCS Algorithm. The architecture consisted of Online and Offline Phase.The Accuracy of prediction is 73%.

Suleyman Salin and Pinar Senkul [7] have used semantic information for web usage mining based recommendation. They have used spade algorithm to generate frequent access sequences. Spade Algorithm is a sequential association rule mining algorithm to generate frequent navigation patterns.

Xin Sui, Suozhu Wang, Zhaowei Li [8] have introduced a framework model of Integration with Semantic Web and Agent Personalized Recommendation System in e-commerce (SWAPRS) is designed based on Semantic Web and agent and web mining technology.

III. Architecture Overview

The architecture for recommending the list of products to the users can be partitioned into two main phases; offline phase and online phase. But these two parts need to work strongly together. Fig.1 shows the architecture. In the offline phase there are two main modules, data pre processing and semantic knowledge base and the main module of the online phase is Recommendation engine.

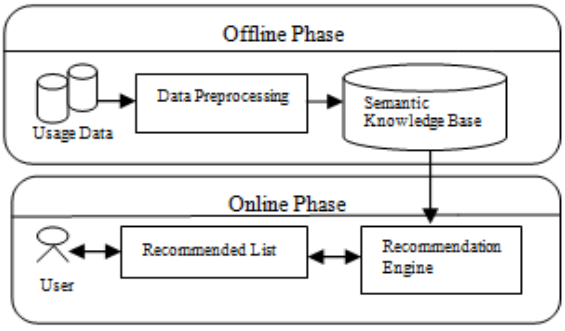

Fig. 1. Architecture of the system

A. Offline phase of the Architecture

This phase consists of two major modules: Data preprocessing and Semantic Knowledge Base of the Products. In this phase we start with the primary Web-Log Preprocessing (Data pretreatment) to extract user session and put the necessary details into the database. After that we integrate the semantic knowledge of the products in to RDF Model. Fig. 2 shows the offline phase.

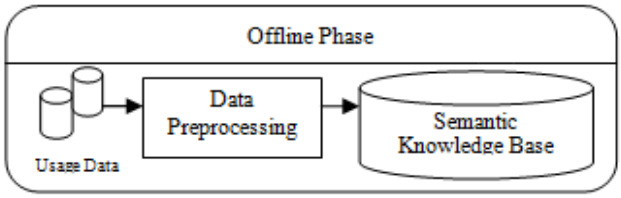

Fig. 2. Offline architecture

1) Data Preprocessing

Data preprocessing in a web usage mining model (Web-Log preprocessing) aims to reformat the original web logs to identify all web access sessions. The Web server usually registers all users' access activities of the website as Web server logs. Due to different server setting parameters, there are many types of web logs (but share basic information like client ip address, URL, HTTP) etc. Generally, several pretreatment tasks (like data cleaning, user differentiation and session identification) need to be done before performing web mining algorithms on the Web server logs.

2) Semantic Knowledge Base.

After the data preprocessing step, we integrate semantic features of the products like price band, brand affinity, rating etc along with the extracted user session information from the web log. We create tables in the database which records the user details and most recent transaction details of the user. To integrate the semantic features of the product purchased by user, we use RDF Model based on the JENA framework.

3) RDF Model

According to the W3C recommendation, RDF is a foundation for processing metadata. It provides interoperability between applications that exchange machine-understandable information on the Web.

RDF documents consist of three types of entities: resources, properties, and statements. Resources may be Web pages, parts or collections of Web pages, or any (real-world) objects which are not directly part of the WWW. In RDF, resources are always addressed by URIs. Properties are specific attributes, characteristics, or relations describing resources. A resource together with a property having a value for that resource forms an RDF statement. A value is a literal, a resource, or another statement. Statements can thus be considered as object–attribute–value triples. The data model underlying RDF is basically a directed labeled graph.

B. Online Phase of the Architecture

During the online phase, when a user arrives at the server, the intelligent agent engine contacts the semantic knowledge base, to know the user's previous transactions. Based on the user's recent history, a list of recommended products will be displayed. If the user is visiting the server for the first time there will be no list of recommended products available until the user purchases any product. All the details about the user

transaction get updated in the underlying knowledge base. Fig. 4 shows the offline phase.

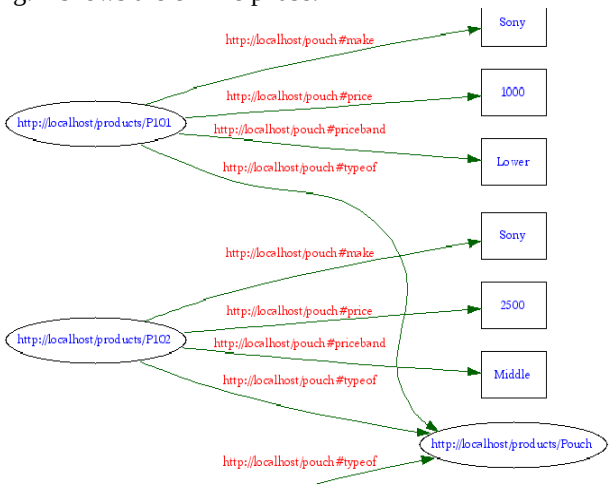

Fig. 3. RDF model of semantic knowledge base

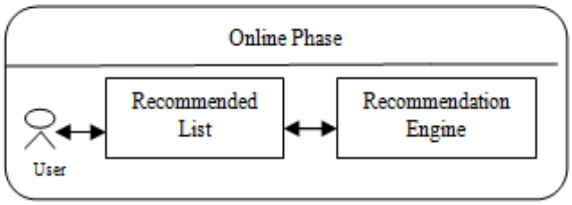

Fig. 4. Online architecture

1) Generating Recommendation

The main objective of recommendation engine is to generate recommendation by using certain filtering parameters like price band, brand affinity, rating etc. The filtering parameters are used to generate precise set of recommended products from the semantic knowledge base. To generate list of recommendation we use Longest Common subsequences algorithm.

IV. Longest Common Subsequence Algorithm

The problem of comparing two sequences A and B to determine their similarity is one of the fundamental problems in pattern matching. One of the basic forms of the problem is to determine the longest common subsequence (LCS) of A and B. The LCS string comparison metric measures the subsequence of maximal length common to both sequences [7].

A. Generating Recommendation by Longest Common Subsequence

Formally, given a sequence α=<1, α2,…, αn > another sequence γ=<β1 β2,…,βn >where β is a subsequence of α if there exists a strictly increasing sequence <j1, j2,…, jn>of indices of α such that for all i=1,2,..,l, we have $\alpha ji = \beta i$. Given two sequence α and β, we say that γ is common subsequence of α and β if γ is a subsequence of both α and β.

We are interested in finding the maximum-length or longest common subsequence (LCS) given two paths or sequence of product information. α=<α1, α2,…, αn >, β =<β1, β2,…, βm >.Pattern search algorithm can be utilized to find user interest products based on the current user activities to predict and recommend user future's request. The pattern search algorithm, Longest Common Subsequences (LCS) is used in the recommendation part of the system.

For example, if the two recommended list of products generated from the underlying knowledge base based on the recently purchased products by the user is Camera/Lens/Pouch and another list Camera/Lens, then the LCS algorithm extracts the common sequence which is recommended to the user. It also recommends the user to buy other products which follow the sequence. Fig. 5 shows how LCS algorithm is applied to the intelligent agent engine.

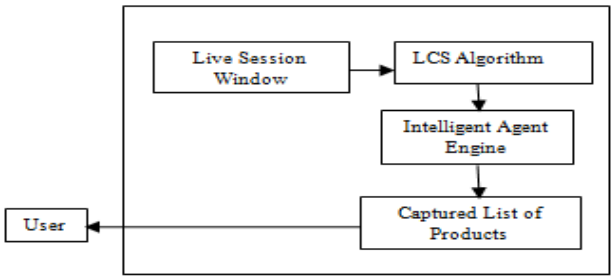

Fig. 5. Applying LCS algorithm

V. Snapshots

The following two figures show the list of the products available and the recommended list of products. We have used JENA Framework to create Semantics of the product in RDF model. We have used Java, J2EE for the front end and MySQL database for the backend. The system is been developed using NetBeans IDE.

Fig. 6. List of products available

Fig. 7. List of recommended products

VI. Conclusion and Future Enhancement

We have created two tier architecture for integrating semantic information with web usage mining. We have used LCS algorithm to generate a list of recommended products to the user.

Future enhancement involves measuring the accuracy of the recommended list to characterize the quality of the results obtained. The EBAY data set will be used for the purpose of performance measure. Our comparison is based on 3 different metrics, namely, accuracy, coverage and F1 measure. Accuracy measures the degree to which the recommendation engine produces accurate recommendations while coverage measures the ability of the recommendation engine to produce all products that are likely to be selected by the user. The $F1$ measure attains its maximum value when both accuracy and coverage are maximized.

$$F1 = \frac{2*\text{Accuracy}*\text{Coverage}}{\text{Accuracy}+\text{Coverage}}$$

REFERENCES

[1] W. T. Yan, M. Jacobsen, and H. Garcia-Molina, Umeshwar, "From user access patterns to dynamic hypertext linking," *Computer Networks and ISDN Systems*, Elsevier, 1996.

[2] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic personalization based on web usage mining," *Communications of the ACM*, ACM, 2000, pp. 142–151.

[3] R. Liu and V. Keselj, "Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests," *Data & Knowledge Engineering*, Elsevier, 2007, pp. 304-330.

[4] R. Baraglia and F. Silvestri, "Dynamic Personalization of Web Sites Without User Intervention," *Communication of the ACM*, 2007, pp. 63-67.

[5] Subhash K. Shinde, and U. V. Kulkarni, "A New Approach For On Line Recommender System in Web Usage Mining," 2008 *International Conference on Advanced Computer Theory and Engineering IEEE* 2008.

[6] Mehrdad Jalali1, Norwati Mustapha, Md. Nasir B Sulaiman, and Ali Mamat, "A Web Usage Mining Approach Based on LCS Algorithm in Online Predicting Recommendation Systems," *12th International Conference Information Visualisation IEEE*, 2008.

[7] Suleyman Salim *et al.*, "Using Semantic Information for web usage mining based recommendation," *International Conference IEEE*, 2009.

[8] Xin Sui, Suozhu Wang, and Zhaowei Li, "Research on the Model of Integration with Semantic Web and Agent Personalized Recommendation System," Proceedings of the 2009 13th International Conference on Computer Supported Cooperative Work in Design.

[9] A. Apostolico, "String editing and longest commonsubsequences," Handbook of Formal Languages, vol. 2. Linear Modeling: Background and Application, Springer Verlag, Berlin, chapter 8, 1997, pp. 361–398.

[10] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, "Introduction to Algorithms," MIT Press, 1990.

[11] V. Dancik, "Expected Length of Longest CommonSubsequences," *PhD thesis*, University of Warwick, 1994.

[12] A. V. Aho, D. S. Hirschberg, and J. D. Ullman, "Bounds on the complexity of the longest common subsequence problem," *J. Assoc. Comput. Mach*, ACM, pp. 1-12, 1976.