Semantic Information and Web based Product Recommendation System – A Novel Approach

Sneha Y.S
Faculty, JSSATE, Bangalore
Research Scholar, Anna University
of Technology Coimbatore, India

G. Mahadevan, PhD.
Prof & HOD, Dept Of CSE,
AMCEC, Bangalore India

ABSTRACT

The World Wide Web is today a perennial source of immense information. There is therefore, a definite demand for automated methods that can locate, identify and retrieve information to cater to the individual's requirements, demands or whims. The internet also creates newer possibilities to organize and recommend information. Web usage mining has become popular in various business areas related with Web site development. As the scale of the Internet is getting larger and larger in recent years, we are forced to spend much time to select necessary information from large amount of web pages created. Traditionally In Web usage mining, commonly visited navigational paths are extracted in terms of Web page addresses from the Web server visit logs, and the patterns are used in various applications including recommendation. But semantic information of the Web page contents is generally not included in Web usage mining. The paper has used OWL technology to add semantics to the existing navigational paths. Results shows that our approach fetched better accuracy than the existing web based approach. This paper presents a framework for integrating semantic information along with the navigational patterns. This paper evaluated the framework and it shows promising results in terms of quality recommendation of products.

KEYWORDS

WUM, Semantic Web, Recommender System, OWL

1 INTRODUCTION

The internet is getting larger and larger in recent years and users are forced to spend lot of precious time to select valuable and necessary information from large amount of web pagers. This has led to Information Overload problem. The users are interested in getting right information at right place at the quickest time as possible. The Explosion of WWW into every body's life, its dramatic growth and the consequent inevitable emergence of ecommerce has led to the development of Recommender System. [17]. Recommender System helps to solve the information overload problem by assisting the users in finding personalized information at the right time. Some of the applications of Recommender System are Personalization, Filtering, and Prediction etc. Recommender System enables ecommerce personalization benefitting business consumer relations by creating a Win-win situation for both consumer and business [18].

The Recommender system can be used to either predict whether a certain user will like a particular item (Prediction Problem,) or to identify a set of N items that would be of interest to the user(top n Recommendations).[19] using information filtering technology. Recommender System is based on two approaches viz Based on user navigation (Implicit) and Based on User ratings (Explicit). This paper deals with user navigation to recommend different products to users. One of the most important applications of Recommender System is Personalization, which deals with learning user's interest, their needs, preferences, tastes, goals etc.

Web Mining is the area of data mining which deals with extraction of the required knowledge from the World Wide Web [7]. Web mining can be broadly classified into three categories depending up on the data to be mined. Web content mining, Web structure mining and Web usage mining. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from web data [21].

In traditional recommender system, web usage mining techniques are used to generate interesting patterns from the past behavior of the user which is present in the web log data. However semantic information is not integrated. Thus if we extract semantics from the web pages which are stored in web server logs and integrate with web usage mining techniques then it is possible to derive better accuracy results.

The number of studies in web usage mining and semantic information is limited. [1,4]. Most of the research has concentrated on user content mining algorithms for generating patterns in terms of semantic information. [1,5,3]. The main work envisages generating user profiles based on graph partitioning algorithm and adding ontological instances in terms of semantic information.

This paper is organized as follows: Back ground and related works is organized in section 2, the Proposed approach is described in section 3 , the performance evaluation of the Proposed approach is described in section 4 followed by Convulsion and future work in section 5.

2. BACK GROUND AND RELATED WORKS

The user's preferences and their navigational behavior of web usage mining system have gone for an extensive change recently. This paper reviews several WUM systems and their architecture for predicting user's choices, preferences etc based on their navigational behavior. This paper also studies how web has gone from its traditional way of searching information to semantic way of searching information.

One of the first WUM systems which came into existence is Analog [6].The Architecture of Analog consist of an offline component and online component. The Offline component is responsible for analyzing the users past behavior and organizing the similar user's behavior into a session cluster. The Online component is responsible for building the active user session and also identifies pages related to each user with list of suggestions and recommendations. The main limitation

of Analog is the usage of geometrical approach while clustering the users. This results in poor scalability. However the architecture is being used in most of the WUM systems.

Perkowitz and Etzioni (who coined the term Data Mining) [7] developed a WUM system which used a partitioning graph theotric approach for mining user logs. This made websites adaptive in nature. They developed a new clustering method called cluster mining which is implemented using Page Gather algorithm.

Web Personalizer is a Two tier architecture developed by Mobasher et al., [8,9] which produces dynamic recommendations. The Recommendations are in the form of hypertext links. Several data mining techniques like clustering, association rule mining and sequential pattern discovery is used to obtain aggregate user profiles in the data pre processing stage of offline phase. The online phase maps the active user sessions with the aggregate usage profiles and computes the set of recommendations.

Liu and Keselj [10] proposed an online recommendations system which calculates user's satisfaction on the list generated recommendations. They used character N grams to combine the contents of web pages and user navigation patterns for building user profiles.

SUGGEST is a WUM system developed by Baraglia and Palmerini [11, 12]. This system also has same architecture as that of Analog i.e. Offline and Online architecture. Offline phase builds the historical data about the user and online phase generates the list of recommendations as per the understanding of user's behavior. They used graph partitioning algorithm to build the active user sessions. Two of the most severe limitations of this architecture are a) Web server pages are quadratic in nature which means more memory is consumed. b) Web sites do not permit to manage dynamically generated pages.

The Semantic web is an extension of the current web in which information is given a well defined meaning and computers will be able to use the information on the web not just present the information.[20].

One of the first research studies on Web Usage mining and semantic web is done by Bettina Berendt, Andreas Hotho and Gerd Stumme [1]. The study consists of two parts. The first part is related to extracting semantics from web page and the second part is to integrate the extracted semantics with web usage mining. A knowledge acquisition method called ONTEX [Ontology exploration] is extensively used in their studies.

The authors [21] in this paper present a semantic based approach to recommender system. In order to overcome the short comings of traditional recommender system, a hybrid recommender system is been proposed. Singular value Decomposition (SVD) and KNN algorithm is being used for generating the features. Naive Bayes Classifier is used for calculating the precision and recall.

The authors in [22] proposed a framework SWAPRS an agent combining the semantic web and personalized recommendation system. The proposed framework establishes semantic model of customer behavior by watching user's past behavior such as what are its shopping habits, preferences etc. Based on this it forms a personalized recommendation set of users and the recommendations are provided. The main drawback of this system is that there is no technique used for clustering the users.

3. SEMANTIC APPROACH FOR WEB USAGE MINING

Mehradad Jalali et al [13] proposed in their studies an on-line Recommendation System using LCS algorithm. For this study they used Graph Partitioning Algorithm for clustering the users and LCS algorithm for generating the list of most common set of recommendations.

Suleyman Salin and Pinar Senkul [14] propose a framework for integrating semantic information with web usage mining. The frequently used navigational patterns are extracted in the form of ontology instances instead of web page addresses and the result is used for generating web page recommendations to the visitor.

Though the authors proposed an online recommender system, the semantic information is not integrated into web usage mining. The authors in [14] used sequential mining for pattern generation by integrating semantic information however there was no technique to group the user. In this research paper the authors have advanced an architecture which adapts the idea of online and offline phase by integrating the semantic information in the offline phase. The list of recommendation is generated by using maximum subsequence problem. The system data flow of the recommender system is shown in Fig 1.

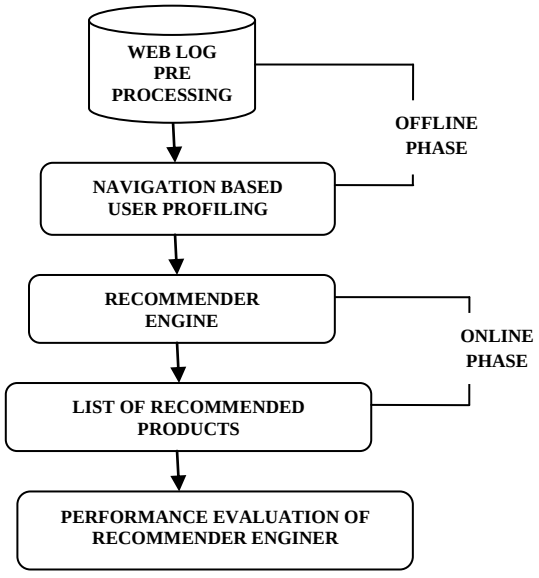

Fig 1 System Data Flow Diagram

The Main architecture consists of two phases Offline Phase and Online phase. The Offline phase and the online phase are responsible for generating list of recommendations. The Main

architecture is shown in Fig 2. The Offline and Online architecture are shown in Fig 3 and Fig 4 Respectively.

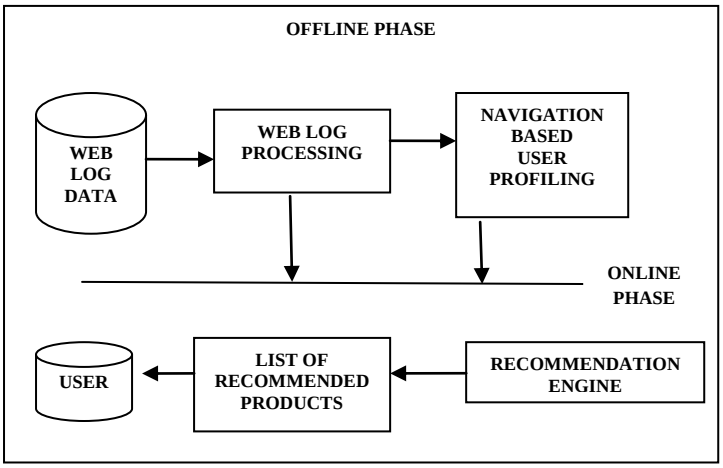

Fig 2 Main Architecture

3.1 *Offline Phase of the Architecture*

This phase consists of 2 modules: Web Log Preprocessing and Navigation Based User Profiling as shown in Fig 2.

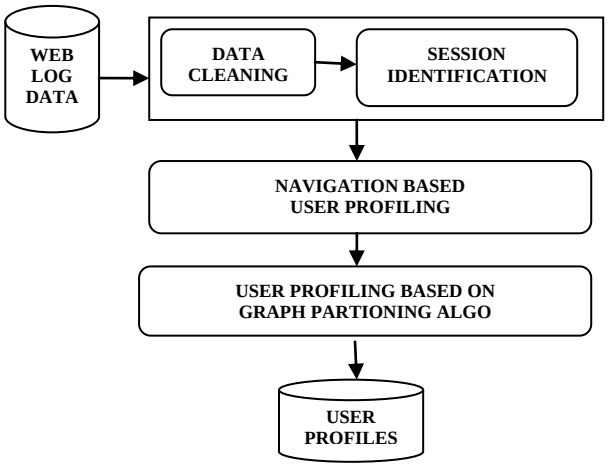

Fig3: OFFLINE PHASE OF THE ARCHITECTURE

3.1.1 *Web Log Preprocessing*

The user's behavior is recorded in web server logs. The basic information about the user such as client id ,address, request time ,requested URL,HTTP etc are recorded in web logs. Some of the pieces of information may be incomplete due to various reasons. This has lead to the cleaning of the data before going into the next phase through web log preprocessing. The main aim of Web Log Processing is to eliminate the incomplete data and format the data to identify the web access sessions. After the preprocessing of the Dataset, this paper maps the Ontology individuals and the requested web page address in terms of classes using OWL format. OWL (Web Ontology Language) is the latest recommendation of W3C [20]. It is the most popular language for creating Ontologies .It is the last technical component in the semantic web architecture. OWL is based on the RDF Schema and expresses much more complex and richer relationships. These relationships help us to recommend products which many users are closely related. The second module offline phase of the architecture is Navigation based User Profiling.

3.1.2 *Navigation Based User Profiling*

After the Web Log preprocessing, we apply Graph Partitioning algorithm in [13] to group users based on their navigation. The algorithm consists of two major factors viz Time Connectivity and Frequency. Time Connectivity describes the order of the visit of the web pages describing the products in a session whereas frequency describes the occurrence of common web pages containing list of products. This Algorithm uses DFS (Dept First Search Algorithm) to find the most correlated pages describing the users. The results obtained after applying Graph Partitioning algorithm is the number of clusters which are stored in the database.

3.2 *Online Phase of the Architecture.*

During the online phase, when a new user arrives, the session to which it belongs is identified and is compared with the buying history of other users. This helps in recommending products to the user. While recommending products to the user, this paper uses the concept if window count. This means that how often these products are recommended. And how many products can be recommended in a single window count. Different values can be given by the user. Depending up on the value of the window count, the list of recommendation is prepared. The online phase of the architecture is shown in Fig 4.

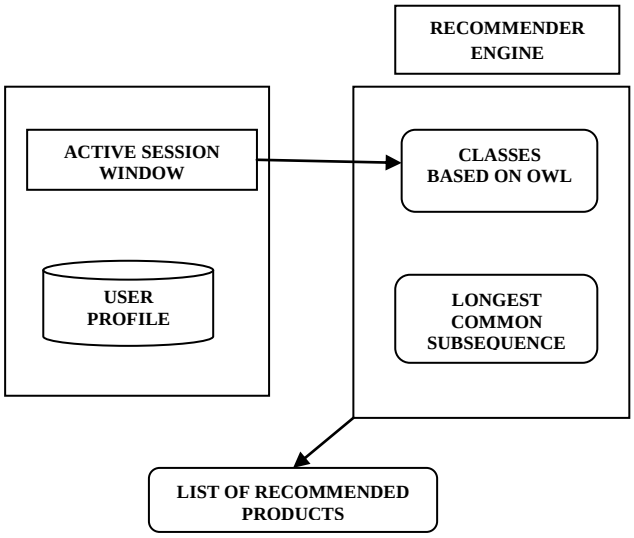

Fig 4 Online Phase Of The Architecture

For the Longest matching pattern we take the longest sub string present in the database. The longest substring is determined using Maximum Subsequence Problem. The Maximum subsequence Problem is linear in nature.

The Maximum Subsequence Problem is given by [16]. Let $X_1, X_2 \text{-----} X_n$ be set of real numbers where X_i corresponds to i^{th} element of the sequence. The Problem is to find contiguous subsequence $X_i, X_{i+1}, \ldots\ldots X_j$ that maximizes $X_i + X_{i+1} \ldots\ldots + X_j$. The algorithm is as follows:

1) The List is searched from right to left for the maximum value of j satisfying $I\,j < I_k$.

2) If there is no such j then add I_k to end of the list.

3) If there is such a j and $R_j > R_k$, then we add I_k to the end of the list.

4) Otherwise we extend the sub sequence I_k to the left to encompass everything up to and including the leftmost I_j and reconsider the newly extended subsequence I_k as in step 1.

4. PERFORMANCE EVALUATION.

This paper has used consumer electronics oriented dataset which consists of 1000 sessions, 150 users, 120 products and 30 classes. The evaluation method of this work is based on the techniques introduced in [15] which define 3 parameters accuracy, precision and F1 Metric. The effectiveness of the recommendation is measured in terms of coverage and precision. 10-fold cross validation is performed for each of the data sets. Each transaction t in the test set is divided into two parts. The first part is the first n items in t for recommendation generation. n is called the window count. The other part, which is denoted as eval, is the remaining portion of t to evaluate the recommendation. Once the recommendation phase produces a set of products, which is denoted as Rec, the set is compared with eval products. This paper splits the session log to 90 % and 10 % at the last. For the users in the 10 % list, this paper finds the recommendation and matching to the recommendation already bought now.

Precision is defined as the proportion of the number of relevant recommendations to the number of all recommendations. Precision measures the accuracy of the recommendations.

$$\text{Precision} = |rec\ I\ eval| \div |rec| \qquad (1)$$

Coverage measures the ability of the recommendation system to produce all the page views that are likely to be visited by the user. In other words, it shows how well the recommendation covers all the pages that the user is likely to visit.

$$\text{Coverage} = |rec\ I\ eval| \div |eval| \qquad (2)$$

This paper calculates precision and coverage with respect to a threshold value. The precision-with-threshold is an extension to the precision measurement. It evaluates the success of the recommendation's precision value with respect to a given threshold values. The value of new precision can be 0 or 1. If the precision is greater than the given precision threshold τ then the value is 1 otherwise 0.

$$\text{Precision with threshold} = \begin{cases} 0 \text{ if precision} < \tau \\ 1 \text{ if precision} \geq \tau \end{cases}$$

The F1 measure attains its maximum value when both precision and coverage are maximized. F1 is the harmonic mean of precision and coverage. The experiments ran for thresholds ranging from 0.1 to 1.0.

The results of the experiments are shown below.

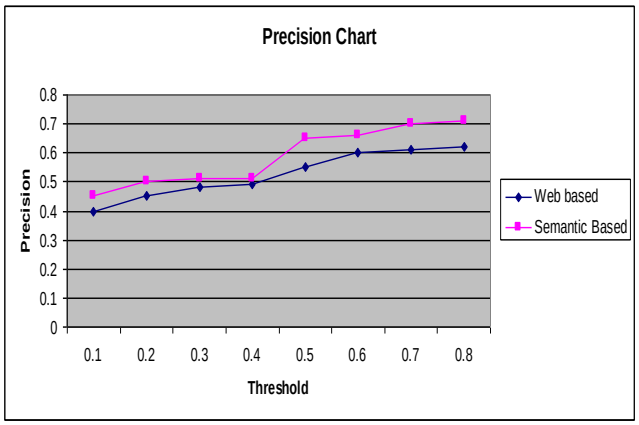

Fig 5 Precision Chart

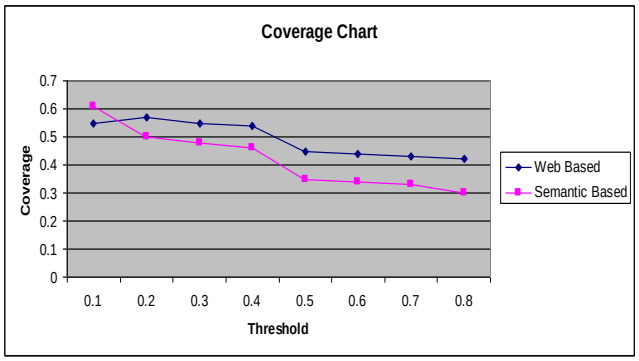

Fig 6 Coverage Chart

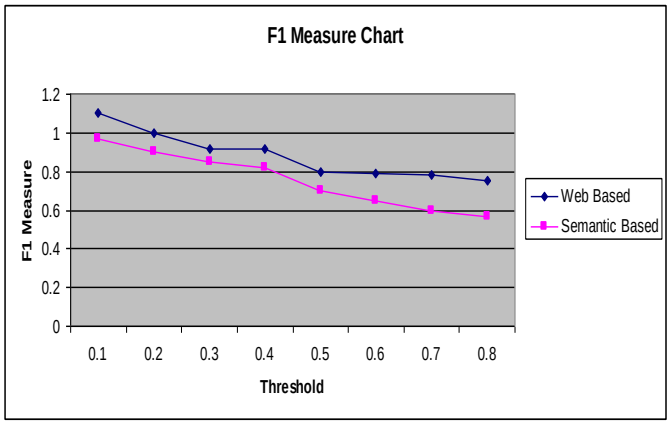

Fig 7 F1 Measure Chart

This research paper compares traditional web based recommendation system with the proposed approach. The performance charts shows the precision of semantic approach is better than that of web based. The reason is that web page based approach just gives the web page recommendation based on navigational pattern of pages , but the proposed approach provides the exact class of product to recommend based on the product bought. Due to this factor, the precision is very sharp in this approach.

5 CONCLUSIONS AND FUTURE WORK

In this paper a novel approach is presented for the Recommender System using web usage mining and semantic web. Semantic web is an emerging area which extends traditional web by adding semantic data to it. This paper has advanced the architecture of the Recommender System by integrating semantic information in terms of OWL instances on web usage mining algorithms.

The Consumer Electronics Oriented Dataset which consists of 1000 sessions, 150 users, 120 products and 30 classes were used for conducting experiments. The paper has used 10 fold cross validation which is performed on the dataset.

The Proposed approach is compared with the traditional web based recommender system. The Graphs show that the proposed approach has attained higher precision than the traditional web based approach.

There are some aspects which can be improved for achieving better results.

The algorithm used for clustering includes 2 factors namely the frequency and time connectivity. For future work there is scope for including influencing factors about users in the recommender system.

There is further scope for integrating semantic information by designing an advanced knowledge base which incorporates ontological instances.

6 REFERENCES

[1] Bettina Berendt, Andreas Hotho and Gerd Stumme, Towards Semantic Web Mining, 2002 In the Proceedings of the First International Semantic Web Conference on The Semantic Web

[2] Gediminas Adomavicius and Er Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions",2005 IEEE Transactions on Knowledge and Data Engineering

[3] Honghua Dai, Bamshad Mobasher, "Integrating Semantic Knowledge with Web Usage Mining for Personalization", Intelligent User Interfaces, IGI Global, 2009.

[4] Bernhard Ganter and Gerd Stumme, "Creation and Merging of Ontology top-levels", Conceptual Structures for knowledge creation and communication, 2003 .

[5] G. Stumme, B. Berendt and A. Hotho, Usage Mining for and on the Semantic Web. Next Generation Data Mining, 2002 in the Proc. NSF Workshop, Baltimore,

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

[7] M. Perkowitz and O. Etzioni, "Towards adaptive Web sites", The International Journal of Computer and Telecommunications Networking,1999

[8] B.Mobasher, R.Cooley, J.Srivastava, ".Automatic personalization based on web usage mining", Communications of the ACM, 2000.

[9] M.Nakagawa, B.Mobasher "A hybrid web personalization model based on site connectivity", ACM Transactions on Internet Technology, 2007.

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

[11] R. Baraglia, F. Silvestri, "Dynamic Personalization of Web Sites Without User Intervention", Communications of the ACM 2007

[12] R.Baraglia, F.Silvestri, 2004 An online recommender system for large Web sites, In the Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence

[13] M. Jalali, N. Mustapha, A. Mamat, Md N. Sulaiman ,OPWUMP, 2008 An architecture for online predicting in WUM-based personalization system , In the proceedings CSICC conference on Advances in Computer Science and Engineering

[14] Suleyman Salin, Pinar Senkul,2009 Using Semantic Information for Web Usage Mining Based Recommendation,In the Proceedings of ISCIS Conference on International Symposium on Computer and Information Sciences,

[15] G. Kowalski, Information Retrieval Systems: Theory and implementation.

[16] W. L. Ruzzo and M. Tompa., 1999, a linear time algorithm for finding all maximal scoring subsequences. In Proceedings of International Conference on Intelligent Systems for Molecular Biology.

[17] Resnick, Paul and Varian Hal ," Recommender system " Communications of the ACM,1997

[18] Tran T et al, 2006, Designing recommender system for e commerce: an integration approach, In the Proceedings of the ICEC, International conference on Electronic commerce.

[19] Oren Etzioni , " The World Wide Web: Quagmire or gold mine?" , Communication of the ACM,1996

[20] Liyang Yu, 2006, Introduction to the Semantic Web and Semantic Web Services.

[21] Loizou, Antonis, Srinandan 2006, Recommender systems for Semantic Web, In ECAI Recommender System Workshop.

[22] Xin Sui, Suozhu Wang, Zhaowei Li, 2009, Research on the model of integration with semantic web and agent personalized recommendation , In the Proceedings of International Conference on computer Supported Cooperative Work In Design.