RESEARCH ARTICLE OPEN ACCESS

Web Page Recommendation Using Web Mining

Modraj Bhavsar[1] Mrs. P. M. Chavan[2]
1 (Student, Department of Computer Engineering & IT, VJTI, Mumbai-19)
2(Professor, Department of Computer Engineering & IT, VJTI, Mumbai-19)

ABSTRACT:
On World Wide Web various kind of content are generated in huge amount, so to give relevant result to user web recommendation become important part of web application. On web different kind of web recommendation are made available to user every day that includes Image, Video, Audio, query suggestion and web page. In this paper we are aiming at providing framework for web page recommendation. 1) First we describe the basics of web mining, types of web mining. 2) Details of each web mining technique.3)We propose the architecture for the personalized web page recommendation.

Keywords: Web mining, web recommendation, web personalization, and User sequence data.

I. INTRODUCTION

Web page recommendations are becoming very popular, and are shown as links to related web page, related image, or popular pages at websites. When user sends request to web server, session is created for the user. During session when user browses a website the list of page that user visits is stored as a session data. This sequence can be organized and stored as web session S = d_1, d_2, d_3, where d_i = page ID of the i^{th} visited page. The main aim of the recommendation system is to predict web page or pages from the user current session data and other user data.

The key feature of the recommendation system is to learn from historic data of the current user as well as other user. The recommendation system decides the domain of the current user from the user's historic data and then predictsthe pages according to the user's domain. Another feature of the recommendation system is to predict web page that is not visited in the user's current session. To achieve these features numbers of issues are evolved.

In the past few years many researchers devoted their work to overcome these issues. Web access sequence (WAS) in Web usage data can be represented approaches based on tree structure and probabilistic model [1].These approaches learn from the training datasets to build the transition links between Web-pages. By using these approaches, given the current visited Web-page (referred to as a state) and *k* previously visited pages (the previous *k* states), the Web-page(s) that will be visited in the next navigation step can be predicted. The performance of these approaches depends on the sizes of training datasets. The bigger the training dataset size is, the higher the prediction accuracy is. However, these approaches make Web-page recommendations solely based on the Web access sequences learnt from the Web usage data. Therefore, the predicted pages are limited within the discovered Web access sequences, i.e., *if a user is visiting a Web-page that is not in the discovered Web access sequence, then these approaches cannot offer any recommendations to this user*. We refer to this problem as "new-page problem" in this study.

Some studies have shown that semantic-enhanced approaches are effective to overcome the new-page problem [2, 3] and have therefore become far more popular. The use of domain knowledge can provide tremendous advantages in Web-page recommender systems [4].

Domain ontology is commonly used to represent the semantics of Web-pages of a website. It has been shown that integrating domain knowledge with Web usage knowledge enhances the performance of recommender systems using ontology-based Web mining techniques [4-6].

Some studies have shown that semantic-enhanced approaches are effective to overcome the new-page problem [2, 3] and have therefore become far more popular. The use of domain knowledge can provide tremendous advantages in Web-page recommender systems [4]. Domain ontology is commonly used to represent the semantics of Web-pages of a website. It has been shown that integrating domain knowledge with Web usage knowledge enhances the performance of recommender systems using ontology-based Web mining techniques [4-6]. Integrating semantic information with Web usage mining achieved higher performance than classic Web usage mining algorithms [5]. However, one of the big challenges that these approaches are facing is the semantic domain knowledge acquisition and representation. How to effectively construct the domain ontology is an ongoing research topic.

This paper presents method to provide better Webpage recommendation based on Web usage data and user's domain knowledge. In this user's session data is collected. Using this bipartite graph is created. This graph is made up of two sets first set is all user sets and second set is of domain set. The edge is drawn from set1 to set2 if the user belongs to some domain of set2. Also there is another bipartite graph with different set, in this graph first set is all domains supported by the system, and second set is collection of all web pages. In this graph also edge is drawn fromset1 to set2 if the web page belongs to some domain.

This paper is structured as follows: Section II briefs the related work. Section III presents the personalized web page recommendation model. VI concludes this paper and highlights some further work.

II. RELATED WORK

Web mining - is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are

1. Web usage mining
2. Web content mining and
3. Web structure mining.

2.1. Web Usage Mining

Web usage mining is the process of extracting useful information from server logs e.g. use Web usage mining is the process of finding out what users are looking for on the Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered:

- Web Server Data: The user logs are collected by the Web server. Typical data includes IP address, page reference and access time.
- Application Server Data: Commercial application servers have significant features to enable e-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.
- Application Level Data: New kinds of events can be defined in an application, and logging can be turned on for them thus generating histories of these specially defined events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the categories above.

2.2. Web structure Mining

Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided into two kinds:
1. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.
2. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

2.3. Web Content Mining

Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. The heterogeneity and the lack of structure that permits much of the ever-expanding information sources on the World Wide Web, such as hypertext documents, makes automated discovery, organization, and search and indexing tools of the Internet and the World Wide Web such as Lycos, Alta Vista, WebCrawler, ALIWEB, MetaCrawler, and others provide some comfort to users, but they do not generally provide structural information nor categorize, filter, or interpret documents. In recent years these factors have prompted researchers to develop more intelligent tools for information retrieval, such as intelligent web agents, as well as to extend database and data mining techniques to provide a higher level of organization for semi-structured data available on the web. The agent-based approach to web mining involves the development of sophisticated AI systems that can act autonomously or semi-autonomously on behalf of a particular user, to discover and organize web-based information.

There are mainly two approaches for web page recommendation.

1) Traditional approach.
2) Semantic based approach.

In tradition approach association rule mining and probabilistic models are commonly used. Models like sequential modeling are effective in the recommendation [2]. Markov models and tree-based structures are goodto show the transition between different web pages in web session[2]. Some surveys [15, 16] have shown that tree-basedalgorithms, particularly Pre-Order Linked WAP-Tree Mining [13], are outstanding in supportingWeb-page recommendation, compared with other sequencemining algorithms.

The semantic basedapproach uses semantic information into Web-page recommendation models. Using ontology of website recommendation system can be improved significantly. For a website domain ontology is useful for classification of the web pages, and this helps to cluster the web pages and searching the web pages. Domain ontology can be obtained by manual or automatic construction approaches. Depending on the domain of interest in the system, we can reuse some existing ontologies or build a new ontology, and then integrate it with Web mining. Web logs in a Web personalization system. In this system, ontology is built with the concepts extracted from the documents, so that the documents can be clustered based on the similarity measure of the ontology concepts. Then, usage data is integrated with the ontology in order to produce semantically enhanced navigational patterns. Subsequently, the system can make recommendations, depending on the input patterns semantically matched with the produced navigational patterns. Liang Wei and Song Lei [18] employ ontology to represent a website's domain knowledge using the concepts and significant terms extracted from documents. They generate online recommendations by semantically matching and searching for frequent pages discovered from the Web usage mining process. This approach achieves higher precision rates, coverage rates and matching rates.

On the other hand, by mapping Web-pages to domain concepts in a particular semantic model, the recommender system can reason what Web-pages are about, and then make more accurate Web-page recommendations [7, 8]. Alternatively, since Web access sequences can be converted into sequences of ontology instances, Web-page recommendation can be made by ontology reasoning [6, 9]. In these studies, the Web usage mining algorithms find the frequent navigation paths in terms of ontology instances rather than normal Web-page sequences. Generally, ontology has helped to organize knowledge bases systematically and allows systems to operate effectively.

III. WEB PAGE RECOMMENDATION SYSTEM ARCHITECTURE

There will be two phases in the whole process – i) offline tasks that includes datapreprocessing and cleaning followed by Pattern mining, ii) online tasks that concern the

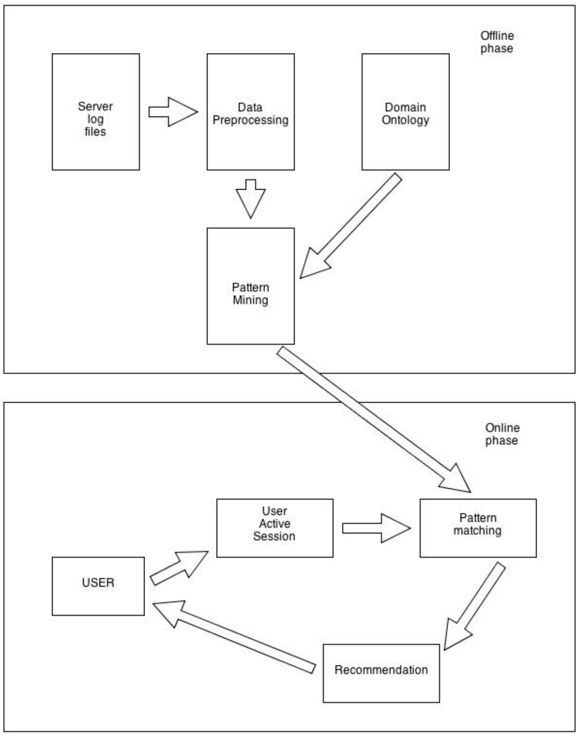

Figure 1: System Architecture

3.1 Data Preprocessing

The preprocessing phase is the first component in the architecture. Web server log file, which is the main source of input, generally contains noisy and irrelevant data. Preprocessing phase consists of data cleaning, user's identification and session identification tasks. During preprocessing Web server log files are pruned to remove irrelevant requests such as non-responded requests and requests made by software agents such as Web crawlers and search engines.

Each Web page is annotated with semantic information during the development of the Website thus showing which ontology class it is an instance of. The cleaned and filtered Web log file is passed to ontology based Web log parser and all the ontology instances represented by theWeb pages are extracted converting the Web log to a sequence of semantic objects. The preprocessing tasks results in aggregate structures such as user transaction file, containing semantic objects where each object is represented as tuple <*page*, *instancei*>, where *page* represents the Web page which contains the object/product, usually an URL address of the page,and *instancei* is an instance of a class c € *C*, from provided ontology *O*, where *i* is an index for anenumeration of the objects in the sequence, from the Web sequence database being mined.

3.2 Pattern Mining

Following the data pre-treatment step, pattern mining is performed on the derived user access sessions. The representative user navigation pattern can be obtained by clustering algorithms. Clustering of user navigation pattern aims to group sessions into clusters based on their common properties. Access sessions that are obtained by the clustering process are actual patterns of Web user activities. User navigation patterns are defined as follows:

Definition 1. A user navigation pattern *np*captures an aggregate view of the behavior of a group of users based on their common interests or information needs. As the results of session clustering, *NP = {np1, np2,. . . , npk}* is used to represent the set of user navigation patterns, in which each *npi*is a subset of P, the set of Web pages.

The process of the clustering takes three steps: are elaborated as follows:

(1) Compute the degree of connectivity between Web pages and create an adjacency matrix.

(2) Create an undirected graph corresponding to the adjacency matrix.

(3) Find connected component in the graph based on graph search algorithm.

Step 1: Compute the degree of connectivity between Web pages and create an adjacency matrix.

For each pair of pages a and b, we compute *W(a, b)*, which is the degree of connectivity between Web pages. A new measurement is proposed for approximating the degree of connectivity for each pair of Web pages in a session, which are Time Connectivity and

Frequency.

Step 2: Create an undirected graph corresponding to the adjacency matrix.

The graph structure can be used to store the weights as an adjacency matrix M where each entry *Mab*contains the value *Wab*computed according to formula in (3). To limit the number of edge in such graph, element of *Mab*whose value is less than the threshold value and is small correlated will be thus discarded. In this study, this threshold is named as *MinFreq*.

Step 3: Find connected component in the graph based on graph search algorithm.

The graph partitioning algorithms divide a graph into k disjoint partitions, such that the partitions are connected and there are a small number of connections between the partitions.

Graph partitioning algorithm is utilized to search for groups of strongly correlated Web pages by partitioning the graph according to its connected components. Depth-first search (DFS) is an algorithm for traversing or searching a graph. Starting from a vertex a, DFS induced by M is applied to search for the connected component reachable from this vertex. Once the component has been found, the algorithm checks if there are any nodes that are not considered in the visit. If so, it means that a previously connected component has been split, and therefore, it needs to be identified. To do this, DFS is applied again by starting from one of the nodes that is not yet visited. In the worst case, when all the URLs are in the same cluster, the cost of this algorithmwill be linear in terms of the number of edges in the complete graph G.

Two main parameters must be accounted for while the algorithm is applied to the undirected graph. Minimum frequency and minimum cluster size are two parameters that significantly affect mining of navigation patterns. MinFreq is a minimum frequency parameter for filtering weights that are below a constant value. The edges of the graph whose values are less than MinFreq are inadequately correlated and are thus not considered by the DFS graph search algorithm. DFS also considers all the connected components that possess the number of nodes greater than a fixed size. Otherwise the rest of components will be considered as insignificant. In this paper, the minimum cluster size is termed as MinClusterSize.

In this study, connected components that have been created based on graph partitioning algorithm are considered as a set of navigation patterns. At the end of this step, the algorithm shows NP = {np1, np2,. . . , npk}, whereby NP is a set of navigation patterns. NP can also be considered as a set of clusters that will further be utilized during the online phase.

The algorithm for navigation pattern mining (clustering) based on graph partitioning algorithm is shown below.

Input:

- Cleaned, filtered, and sessionized Log file.
- MinFreq.
- MinClusterSize.

Output:

- List of Clusters C

```
L[p] = P; //Assign all URL's to a list of web //pages
foreach (Pi, Pj) € L[ p ] do //for all pair of //web pages
    M (i, j) = WeightFormula (Pi, Pj); //computing the weight based on formula (3)
    Edge (i, j) = M (i, j) ;
end for
//There is an undirected graph (E, V)
forall Edge (u, v) € Graph (E, V) do //removing all edges that its weight is below //than MinFreq
    ifEdge (u, v) <MinFreqthen
        remove (Edge(u, v)) ;
    end if
end for
forall vertices (u) € Graph (E, V) do
    Cluster [i] = DFS (u) ;//doing the DFS //algorithm
    ifcluster[i] <MinClusterSize//remove the //cluster that its size is below than //MinClusterSize
        remove (Cluster[i]);
```

```
end if
i = i+1
end for
return (Cluster) ;
```

3.3 Online Recommendation Phase

The aim of a recommender system is to determine which Web pages are more likely to be accessed by the user in the future. In this phase active user's navigation history is compared with the discovered Sequential Association rules in order to recommend a new page or pages to the user in real time. Generally not all the items in the active session path are taken into account while making a recommendation. A very earlier page that the user visited is less likely to affect the next page since users generally make the decision about what to click by the most recent pages. Therefore the concept of window count is introduced. Window count parameter '*n'* defines the maximum number of previous page visits to be used while recommending a new page. Since the association rules are in the form of ontology individuals, the user's navigational history is converted into the sequence of ontology instances. Then the semantic rich association rules and user navigation history are joined in order to produce recommendations.

In the recommendation phase, in the first instance, the most recently navigated item is taken as the search pattern. All the semantic–rich association rules are scanned and the association rules whose antecedent part is equal to the search pattern are added to the recommendation set. This step iterates window count times and at each iteration, the search pattern is extended by one item. The recommendation set constitutes the set of semantic rich association rules sorted in the decreasing order of their confidence. After constructing the recommendation set, the page recommendation commences. Semantic distance between objects is taken into consideration to solve the ambiguity problem. For instance consider the following two semantic rich association rules and

AB ---> C
AB--- >D

where A, B, C, D are semantic objects. If semantic distance (B,D) < semantic distance(B,C) meaning that D is semantically closer to B than C is, then recommendation engine will prefer D over C and the page(s) representing product D will be recommended. Such capability is not provided by regular association rules. The consequent part of the rule contains ontology individuals; therefore the instances should be converted to the real Web objects. The Web pages for the Web objects present in the recommendation set are recommended.

IV. FUTURE WORK

There are a number of aspects that merit further improvement by the system. We can take into account the semantic knowledge about underlying domain to improve the quality of the recommendations. Integrating semantic Web and Web usage mining can help in achieving best recommendations from the dynamic and huge Web sites. The recommendations will be much more relevant, since they will be some relation to each other; instead of just following the navigation patterns.

REFERENCES

[1] B. Liu, B. Mobasher, and O. Nasraoui, "*Web Usage Mining*," in Web Data Mining: Exploring Hyperlinks, Contents, and UsageData, B. Liu, Ed.: Springer-Verlag Berlin Heidelberg, 2011, pp.527-603.

[2] B. Mobasher, "*Data Mining for Web Personalization*," in TheAdaptive Web. vol. 4321, P. Brusilovsky, A. Kobsa, and W. Nejdl,Eds.: Springer-Verlag Berlin, Heidelberg, 2007, pp. 90-135.

[3] G. Stumme, A. Hotho, and B. Berendt, "*Usage Mining for and on the Semantic Web*," AAAI/MIT Press, 2004, pp. 461-480.

[4] H. Dai and B. Mobasher, "*Integrating Semantic Knowledge with Web Usage Mining for Personalization*," in Web Mining:Applications and Techniques, A. Scime, Ed. Hershey, PA, USA: IGI Global, 2005, pp. 276 - 306.

[5] S. A. Rios and J. D. Velasquez, "*SemanticWeb Usage Mining by a Concept-Based Approach for Off-line Web Site Enhancements*," in Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on, 2008, pp. 234-241.

[6] S. Salin and P. Senkul, "*Using Semantic Information for Web Usage Mining based Recommendation*," in 24th International Symposium on Computer and Information Sciences, 2009., 2009, pp. 236-241.

[7] A. Bose, K. Beemanapalli, J. Srivastava, and S. Sahar, "*Incorporating Concept Hierarchies into Usage Mining Based Recommendations*," in Proceedings of the 8th Knowledgediscovery on the web international conference on Advances in webmining and web usage analysis Philadelphia, PA, USA: Springer- Verlag, 2007, pp. 110-126.

[8] N. R. Mabroukeh and C. I. Ezeife, "*Semantic-Rich Markov Models for Web Prefetching*," in 2009 IEEE InternationalConference on Data Mining

Workshops Miami, Florida, USA, 2009, pp. 465-470.

[9] M. O'Mahony, N. Hurley, N. Kushmerick, and G. Silvestre, "*Collaborative recommendation: A robustness analysis,*" ACMTransactions on Internet Technology, vol. 4, pp. 344-377, 2004.

[10] G. Stumme, A. Hotho, and B.Berendt, "*Semantic Web Mining: State of the art and future directions,*" Web Semantics: Science,Services and Agents on the World Wide Web, vol. 4, pp. 124-143, 2006.

[11] B. Zhou, S. C. Hui, and A. C. M. Fong, "*CS-Mine: An Efficient WAP-Tree Mining for Web Access Patterns,*" in Advanced WebTechnologies and Applications. vol. 3007: Springer Berlin / Heidelberg, 2004, pp. 523-532.

[12] J. Borges and M. Levene, "*Generating Dynamic Higher-Order Markov Models in Web Usage Mining,*" in Knowledge Discoveryin Databases: PKDD 2005. vol. 3721: Springer Berlin / Heidelberg, 2005, pp. 34-45.

[13] C. I. Ezeife and Y. Lu, "*Mining Web Log Sequential Patterns with Position Coded Pre-Order Linked WAP-Tree,*" Data Mining andKnowledge Discovery, vol. 10, pp. 5-38, 2005.

[14] B. Zhou, "*Intelligent Web Usage Mining,*" Nanyang Technological University 2004.

[15] C. Ezeife and Y. Liu, "*Fast Incremental Mining of Web Sequential Patterns with PLWAP Tree,"* Data Mining and KnowledgeDiscovery, vol. 19, pp. 376-416, 2009.

[16] T. T. S. Nguyen, H. Lu, T. P. Tran, and J. Lu, "*Investigation of Sequential Pattern Mining Techniques for Web Recommendation,*" International Journal of Information and Decision Sciences(IJIDS), pp. x-x, 2012.

[17] S. T. T. Nguyen, "*Efficient Web Usage Mining Process for Sequential Patterns,*" in Proceedings of the 11th InternationalConference on Information Integration and Web-based Applications & Services, Kuala Lumpur, Malaysia 2009, pp. 465-469.

[18] L. Wei and S. Lei, "*Integrated Recommender Systems Based on Ontology and Usage Mining,*" in Active Media Technology. vol. 5820, J. Liu, J. Wu, Y. Yao, and T. Nishida, Eds.: Springer-VerlagBerlin Heidelberg, 2009, pp. 114–125.