IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661 Volume 2, Issue 5 (July-Aug. 2012), PP 34-37
www.iosrjournals.org

HYBRID WEB MINING FRAMEWORK

Prof. (Mrs) Manisha R. Patil [1] and Mrs. Madhuri D. Patil [2]
[1] *Professor, Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Pune, India*
[2]*Student, Department of Computer Engineering Smt. Kashibai Navale College of Engineering, Pune, India*

ABSTRACT: *IN THIS PAPER WE INTRODUCE A FRAMEWORK FOR WEB PERSONALIZATION THAT COMBINES THREE WEB DATA MINING TECHNIQUES TO PROVIDE RECOMMENDATIONS TO USER MORE ACCURATELY AND IT USES THREE ALGORITHMS ON THE RESULTS OF PREPROCESSING OF WEB DATA'S K-MEANS ALGORITHM USES VECTOR SPACE MODEL TO REPRESENT DOCUMENT AND THE SYSTEM GENERATE THE RECOMMENDATION BASED ON PATTERN ANALYSIS RANKING.*

KEYWORDS: *Web Mining, Personalization, Usage Mining, Structure Mining, Association Rule Mining, Content Clustering and Hybrid Recommendation.*

I. INTRODUCTION

With the dramatically quick and explosive growth of information available over the Internet, World Wide Web has become a powerful platform to store, disseminate and retrieve information as well as mine useful knowledge. A user interacts with the Web; there is a wide diversity of user's content they prefer and the navigational patterns they prefer, which results in needing different contents and presentations of information. To improve the Internet service quality and increase the user click rate on websites, it is necessary for a Web developer or designer to know what the user really need to do, predict which pages the user is interested in, and present the customized Web pages to the user by combining user navigational pattern data, web content data and these with website structure data [1], [2], [3].

Click stream data is a principle and rich source of data for analyzing user behavior and interests to making personalized recommendations, but the volume and noisy nature of this data makes it difficult to identify and mine behavioural patterns correctly. Recently, web mining techniques have been widely used for discovering interesting and useful patterns from the World Wide Web. In general, there are three research areas in web data mining, based on mining goals, which determine the part of web to be mined: web structure mining, web content mining and web usage mining [4,5]. Web structure mining is closely related to analyzing hyperlinks and link structure on the web for information retrieval and knowledge discovery. Web structure mining can be used by search engines to rank the relevancy between websites classifying them according to their similarity and relationship between them [6]. Google search engine, for instance, is based on Page Rank algorithm [7], which states that the relevance of a page increases with the number of hyperlinks to it from other pages, and in particular of other relevant pages. Web structure mining is used for identifying "Authorities", which are web pages that are pointed to by a large set of other web pages that make them candidates of good sources of information [8]. Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the web server logs is web usage data mining. The web server logs contains: The domain name (or IP address) of the request; the name of the user who generated the request; the date and time of the request; the method of the request; the name of the file requested; the result of the request (success, failure, error, etc.); the size of the data sent back; the URL of the referring page; the identification of the client agent; and a cookie, a sting of data generated by an application and exchanged between the client and the server. By mining Web usage data, more complete knowledge about Web usage can be obtained. Web usage mining has been an important technology for understanding users' behaviours on the Web [9].

The main contribution of this paper is a hybrid recommendation framework that first discovers weighted association rules from usage data and has content clustering on content data and then combines structure data with these two to recommend pages accurately.

II. BACKGROUND

Personalization consists of three main phases including pre-processing click stream data, pattern discovery and recommendation [10]. The data pre-processing phase transforms raw web data files into suitable data that

can be used by data mining tasks. A variety of data mining techniques can be applied to this formatted data in the pattern discovery phase, such as clustering, association rule mining, and sequential pattern discovery. The results of the mining phase are transformed into aggregate profiles, suitable for use in the recommendation phase. The recommendation engine considers the active user session in conjunction with the discovered patterns to provide personalized content. Web usage mining techniques have been widely applied for discovering interesting and frequent user navigation patterns from web server logs. association rule mining discover different access patterns from web logs that can be modelled and used to offer a personalized and proactive view of the web services to users. At the same time, web content mining approaches have also been investigated and implemented for extracting knowledge from the contents of websites. For example, the clustering and classification of web pages are typical application of content mining techniques [11] and also at the same time web structure mining approaches are also used. For example, page ranking and link analysis are the techniques to have patter discovery of web structure data. The majority of the proposed personalization architectures focus on the use of usage data [12], [13], and [14] and only a few efforts also incorporate knowledge associated with the content [15] or the structure [16] of the web site. As noted in [17], usage-based personalization can be problematic either when there is not enough usage data in order to extract patterns related to certain categories, or when the site content changes and new pages are added but are not yet included in the web log. A few hybrid web recommender systems have been proposed in the literature [17]. More recently, systems that take advantage of a combination of content, usage and even structural information of the websites have been introduced and shown superior results in the web page recommendation problem. Association Rule (AR) mining can lead to higher recommendation precision, and are easy to scale to large datasets, in recently previous study Weight is incorporate into the AR model[19] but it is not combining three patter analysis and generating recommendation score based on three pattern discoveries generated from pattern discovery phase.

III. NEW PERSONALIZED FRAMEWORK

Our system is composed of four modules: an pre-processing which can be subdivided into three part-first to have usage data pre-processing second pre-processes content data and third will pre-processes website structure and output of these three sub units are taken for pattern discovery and then pattern analysis is performed and finally the recommendation engine generates recommendation based on recommendation score. Fig. 1 depicts the general architecture of our system. Entries in a web server log are used to identify users and visit sessions, while web pages or resources in the site are pre-processed and structured data is combined to these two data after pre-processing.

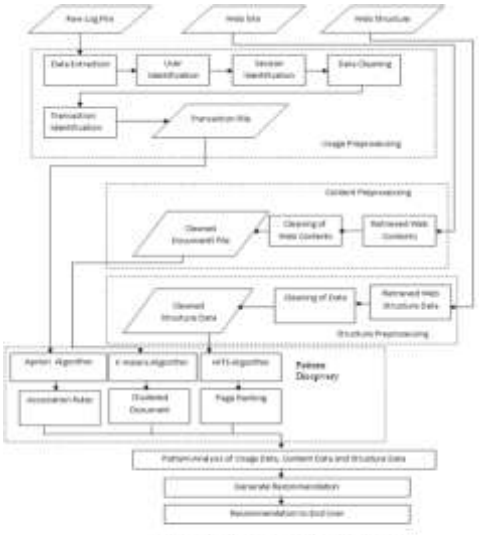

Figure 1 : Hybrid Web Personalization Framework

3.1 Data Pre-Processing

Before using data for pattern discovery it need to be cleaned to get data in specific format. Pre-processing converts the raw data into the data abstractions necessary for pattern discovery. The purpose of data pre-processing is to improve data quality and increase mining accuracy. Pre-processing consists of field extraction, data cleansing. This phase is probably the most complex and ungrateful step of the overall process. This system only describe it shortly and say that its main task is to ”clean” the raw web data such as web logs,

content data, structure data and insert the processed data into a relational database, in order to make it appropriate to apply the data mining techniques in the second phase of the process.
So the main steps of this phase are:
1) Extract the web logs that collect the data in the web server.
2) Clean the web logs and remove the redundant information.
3) Parse the data and put it in a relational database or a data warehouse and data is reduced to be used in pattern analysis to create summary reports.

3.2 Pattern Discovery

The Pattern Discovery Phase is the key component of the Web mining. Pattern discovery converge the algorithms and techniques from several research areas, such as data mining, machine learning, statistics, and pattern recognition, etc. applied to the Web domain and to the available data [19].

3.3 Content Clustering

At the final stage of pre-processing of content data we have m web pages that have used to mining content of these pages and integrating them with usage and structure patterns. Through-out this paper we will use the symbols m, n, and k to denote the number of documents, the number of terms, and the number of clusters, respectively. We will use the symbol S to denote the set of n documents that we want to
cluster, C_j, C2, • • • , C_k to denote each one of the k clusters, and n1 n2, . . . , nk to denote the sizes of the corresponding clusters.

The K-means clustering algorithm that is used in this paper use the vector-space model to represent each document. In this model, each document d is considered to be a vector in the term-space. In particular, we employed the *tf-idf* term weighting model, in which each document can be represented as:

$$(tf1 \log(n/df1), tf2 \log(n/df2),\ldots, tfm \log(n/dfm)$$

Where tf_i is the frequency of the *i* th term in the document and df_i is the number of documents that contain the *i* th term. To account for documents of different lengths, the length of each document vector is normalized so that it is of unit length ($\| dtf_i df \| = 1$).
In the vector-space model, the cosine similarity is the most commonly used method to compute the similarity between two documents d_i and d_j as (1).

$$cos(d_i d_j) = \sum_{k=1}^{m} d_i^k d_j^k \quad (1)$$

Assume that we know in advance the number of clusters that the algorithm should produce. The best known approach that is based on partitioning is k-means clustering, a simple and efficient algorithm used by statisticians for decades. The idea is to represent the cluster by the centroid of the documents that belong to that cluster (the centroid of cluster C is defined as. The cluster membership is determined by finding the most similar cluster centroid for each document. After clustering done, similar pages are assigned to same cluster that can be used in recommendation process.

3.4 Page Ranking

Finally, by employing the HITS algorithm on structure data system generate ranked pages. In HITS concept, Kleinberg identifies two kinds of pages from the Web hyperlink structure: authorities (pages with good sources of content) and hubs (pages with good sources of links). For a given query, HITS will find authorities and hubs. According to Kleinberg, "Hubs and authorities exhibit what could be called a mutually reinforcing relationship: a good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs".

IV. GENERATE RECOMMENDATIONS

The recommendation engine is the online component of a personalization system in order to determine which items are to be recommended, a recommendation score is computed or each page p as in (8).Two factors are used in determining this recommendation score: the overall matching score of the active session(S) to the weighted rules as a whole, and the weighted confidence of the rule.

$$Recommendation\ Score\ (p) = Similarity\ (S, R_L)\ x\ Confidence\ (R_L \rightarrow P)$$

We choose the highest recommendation score as the recommendation to the active session.

V. RECOMMENDATION TO END USER

Based on evaluation and comparison the recommended pages are displayed to end user.

VI. CONCLUSIONS

In this paper, a new web page recommendation framework is proposed. First, users' navigational patterns are extracted from web usage data simultaneously web content data and web structure data is also taken after pre-processing and pattern discovery is performed on these data's and based on the pattern discovery the recommendations are generated. Proposed framework is combining three mining techniques so it is advantageous as compare to the previous hybrid recommendation frameworks.

REFERENCES

Conferences

[1] Agrawal R. and Srikant R. (2000). Privacy preserving data mining, In Proc. of the ACM SIGMOD Conference on Management of Data, Dallas, Texas, 439-450.

[2] Berners-Lee J, Hendler J, Lassila O (2001) The Semantic Web. Scientific American, vol. 184, pp34-43.

[3] Berendt B., Bamshad M, Spiliopoulou M., and Wiltshire J. (2001). Measuring the accuracy of sessionizers for web usage analysis, In Workshop on Web Mining, at the First SIAM International Conference on Data Mining, 7-14.

[4] Srivastava, et aI. , Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, 1(2) 2000,p. 12-23 (3).

[5] R. Kosala and H. Blockeel. Web Mining Research: A Survey, ACM SIGKDD Explorations Newsletter, June 2000, Volume 2 Issue 1.

[6] Kosala, R., and Blockeel, H., (2000). Web Mining Research: A Survey, ACM 2(1):1-15.

[7] Brin, S., and Page, L. (1998). The Anatomy of a Large- Scale Hypertextual Web Search Engine, *Proceedings of the 7th International World Wide Web Conference*, Elsevier Science, New York, 107-117.

[8] Desikan, P., Srivastava, J., Kumar, V., and Tan, P.N. (2002). *Hyperlink Analysis: Techniques and Applications*, Technical Report (TR 2002-0152), Army High Performance Computing Center.

[9] Li Haigang Yin wanling "Study of Application of Web Mining Techniques in E-Business" IEEE Conference , 2006

[10] B. Mobasher, H. Dai, T. Luo, M. Nakagawa. Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. In Data Mining and Knowledge Discovery, Kluwer Publishing, Vol. 6, No. I, pp. 61-82, January 2002.

[11] D. Shen, Y. Cong, J.-T Sun, Y.-c. Lu, Studies on Chinese web page classification, in: Proceedings of the 2003 International Conference on Machine Learning and Cybernetics, 1(2003), pp. 23-27.(20)

[12] B. Mobasher, R. Cooley, 1. Srivastava, Creating Adaptive Web Sites Through Usage-Based Clustering of URLs, in Proc. of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99),1999(24)

[13] B. Mobasher, R. Cooley, 1. Srivastava, Automatic Personalization Based on Web Usage Mining, Communications of the ACM, August 43(8),2000,142-151(25).

[14] M. Spiliopoulou, Web Usage Mining for Web Site Evaluation, Communications of the ACM, 43(8), 2000, pp: 127-134.(29)

[15] H. 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.(42)

[16] M. Eirinaki, M.Vazirgiannis, L Varlamis, SEWeP: Using Site Semantics and Taxonomy to Enhance the Web Personalization Process, in Proc of the 9th SIGKDD Cont: 2003.(37)

[17] B. Mobasher, H. Dai, T. Luo, Y Sung, 1. Zhu, Integrating Web Usage and Content Mining for More Effective Personalization, in Proc. of the International Conference on E-Commerce and Web Technologies (ECWeb2000), 2000.(34)

[18] Samira Khonsha, Mohammad Hadi Sadreddini "Hybrid Web Personalization Framework" in 978-1-61284-486-2/111©2011 IEEE.

[19] José Roberto de Freitas Boullosa. "An Architecture for Web Usage Mining".

[20] Mrs Surekha R.Deshmukh, Dr. D.J. Doke, Dr. Y.P. Nerkar, "Optimal Generation Scheduling with Purchase of Power from Market and Grid ," IEEE Conference 'TENCON 2009', 23-26 Nov. 2009, Singapore.

[21] R. Agrawal, R. Srikant, Fast Algorithms for Mining Association Rules in Large Databases, In Proceedings of the 20th International Conference on Very Large Data Bases VLDB'94,Santiago, Chile, 1994, pp. 487-499.(60)

Authors Biography:

*

Prof. (Mrs) Manisha R. Patil, Professor in Computer Engineering Department of Smt. Kashibai Navale College of Engineering, Vadgaon (Bk), Pune and having 14 yrs of teaching experience. Her area of interest is Data Mining.

*

Mrs. Madhuri D. Patil, Student of Computer Engineering Department in Smt. Kashibai Navale College of Engineering, Vadgaon (Bk), Pune, pursuing M. E. in Computer Engineering and her area of interest is Web Mining.