Volume 3, Issue 10, October – 2018 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165

A Schematic Data Mining Approach for Web Page Recommendation Systems

Manikandan .R
Research Scholar
Anna University

Ramesh R
Assistant professor
Sri Krishna Arts and Science College, Coimbatore

Abstract:- A Novel approach is proposed for a framework based on the data mining techniques for online page recommendation . A recommender system is a software agent that helps in automatically detecting the information which suits an individual need. The recommendation can be made in term of two aspects namely the content based and the collaborative filtering method. With the proper usage of data mining algorithms, the filtration process can be completed before the actual recommendation. This will result in the improvisation of response time.

I. INTRODUCTION

As the data on the web is increasing rapidly, information is over loaded, particularly in web, to overcome this problem[7] many information retrieval systems has been deployed in the recent times to help the users getting the information they need from the Web. The most common approach is by using the key word search which is widely adopted by many search engines. Te lack of enough knowledge on the retrieval process, it becomes tedious and difficult to obtain the desired information. The task of customization of the output according to the user preferences seems to be a great overhead. An alternate approach would be to recommend the users in the system. Recent research have developed a number of recommendation system across different domains which includes news, papers, articles, e-commerce [8,9,10,12]. The two important ways to filter the recommendations are 1) content based filtering and 2) Collaborative filtering. The major flaw occurs when most of the filtering are taking place at the recommendation level. On order to recommend web pages to a particular user, the Nearest Neighbor clustering is normally used. [9] The major issue with the NN clustering is the scalability. i.e., the response time gradually increases with the increase in number of users. The cold-start problem is also a major issue. A new approach called content based filtering is hence considered as an another means of filtering[1][2].

A new approach is presented in this paper so as to combine the content based and collaborative methods with data mining techniques. Both the textual analysis and the user profile creation is done during the recommendation process. While the user browsing patterns are mined to achieve collaborative filtering by applying rule mining algorithms with Predefined constraints. The precision and recall are enhanced in terms of prediction and thus avoiding non repetitive web pages. The Experimental results have proven that our method has a potential to reduce the access time when compared to Markov model [5]. The efficiency is improved by applying the association rule mining for the filtering process as it helps in generating the recommended list of pages before the actual process of recommendation. The response time is hence enhanced. Another advantage of using Data mining techniques is that it helps in data reduction and can handle large volume of data hence fit for large websites and domains. This framework is then applied to an e-learning environment which automatically generates a list of on-line pages based on a individual preference. The typical function of a query based recommendation system is out smarted. The user profile creation is an important component as it stands as a variation from normal key word search typed recommendation. The system is allowed to monitor the user profiles and is made to learn from the past preferences and also the new preferences are updated to achieve better results. A database is maintained for the smooth information access and retrieval. All the related components such as text, links, pages and user profiles are stored in the data base. The front end web page is created and linked with DB for making the user interface better. The traditional methods have no way in customizing the results.

The information retrieval is considering an new approach based on the recommendation of the users in terms of reviews, ratings and likes. The recommender systems have been proposed in multi domain that includes e-commerce, health care, movies, hotels, etc. [8,9, 10] . The recommender system is constructed with consideration of two main aspects first is by the content analysis and the second is through the references of users. The major issue with many of the recommender systems is that it considers one filtering at one time. The nearest neighbor algorithm is mainly used to cluster and compare the users' ratings that are closely in resemblance to the user profile [9]. The major draw back of the nearest neighbor algorithm is the scalability I.e. the time for execution. The second approach also suffers from cold-start problem. The alternate is achieved through the content based filtering which focuses on the classification of textual content.

In this paper a new framework is proposed that combines both content-based and collaborative filtering and utilizes data mining techniques. Both the text analysis and the mining of user access patterns are carried out by applying association rule mining. This is done in order to predict the web pages to include non consecutive pages which results in the performance in terms of precision and re-call. The experimental results also shows that the response time is reduced when compared with Markov model[5,6]. The application of data mining have resulted in increased efficiency since the recommended list of information is generated before the actual process of recommendation. The additional advantage of using the data mining technique is that with the use of data reduction which makes the system scalable for large data sets and domains. The proposed framework is then applied to the e-learning website of a university which automatically comes out with a list of recommended pages on a user preference.

The core concept of the work is information filtering based on the specification of the user. The proposed system considerably increased the potential of the traditional query bases IR systems by auto generation of web pages based on a particular users interest. User profile creation and monitoring is an important aspect in the proposed system as it studies the users browsing patters and also adds new information as and when the user needs or browsing pattern changes and thus results in increased efficiency. To have an easier data access and retrieval, a data base is designed and mounted on the back end of a web user interface. This web user interface is hosted and works on HTTP protocol. The Data base is designed in such a way that it stores all the web components such as textual content, link structure and recommended pages. The remainder of this paper is organized as follows. In the next section, the data mining framework is presents followed by the prototype explanation of web page recommendation with implementation. The conclusion is given in the Section 4

II. THE PROPOSED RECOMMENDER SYSTEM FRAMEWORK

Data mining is the trivial task of retrieving knowledge from a data base which is relatively large[3,4]. The data mining techniques are applied here in the context of IR as we have used IR tools in the proposed frame work. The major classification IR when data mining is applied is content-based filtering and collaborative filtering. Association rule mining is applied as a part of content based filtering where the key word matrix is used as the data set. Then the collaborative filtering is achieved through the data mining of users' browsing patterns. The overall design and implementation consists of the following;

- *Data Collection*

The initial step which collects data that is suitable for applying data mining. Three components are considered and are textual content, the structure of link and web server logs

- *data pre-processing*

This step is done to clean the data and transform the same into suitable formats for the application of data mining algorithms. The data reduction and some of the selection techniques are followed to improve the efficiency of the mining algorithms.

- *IR using data mining*

This is the major part of the where the analysis of the data is made and applied to generate the useful and recommended web pages.

- *Database Design and Implementation*

This step is done to design a data base to produce ease of access and retrieval of data. All the aspects such as text, links and recommended pages are stored in the data base.

User Interface Design and Implementation: This acts as a bridge between the user and the RS. This step is intended to design and develop web pages that connect with the data base on the back end. The request and service are achieved through HTTP protocol. This interface is mainly used to personalize the web pages that are intended for a particular user and also logs the events to the data base as and when a new pattern is found.

III. A PROTOTYPE

This Web-page recommender system prototype focuses on the university website as the main scope. The site provides information about the university, courses and the services it has on own. The users are categorized as general, students, teachers, parents and admin who are interested in finding the information about the university. The Web site has the main page which can be accessed through a URL. The information can currently be viewed through traditional query based search or through hyperlinks within the page. The proposes system is designed with a focus of recommending pages that are of users interest

- *Data collection and preprocessing*

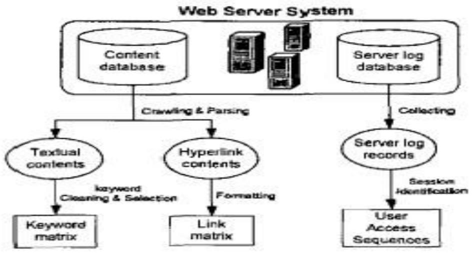


Fig 1:- Data Collection and Pre-processing

Figure 1 illustrates the data collection and preprocessing steps. There are two data bases considered here the first one the content base and the server log base. A web crawler program is used to collect the raw data from the web site. The server log is recorded using a java program . The number of unique web pages that was recorded as output was 21,650. the text and the hyper links are collected separately from the crawling process. The data cleaning is done using the stop word removal and non discriminate words are removed. Then the document-frequency key word selection is applied and a set of key words are selected. [14]. Then the matrix is constructed with no of documents and no of key words. The embedded hyper links are extracted and the link structure is obtained. In the internet, each hyperlink represents a individual web page and hence a matrix of 21650 documents by 21650 links are constructed. The space complexity is taken care by choosing a suitable data structure.

The other data set is collected from the user logs. Each request from a client to the server is recorded and it is termed as a transaction. The recording is done in accordance with the traffic that prevailed in the week days and week ends and thus the monitoring and log collection was made for a complete week cycle. Then the web mining process is applied. The records from both the servers are combined using a Merge sort algorithm[11] in ascending order of time. The scope of the work is for HTML and hence the other multimedia data are removed using the cleaning process. This can be done using a separate program to find out the extension of a file and classifying and removing the multimedia files such as image, audio, video ,etc. The user access sequence is carried out using the session identification. The IP address are used to identify distinguished users. The browsing path or a traverse path is obtained by continuous recording of the session of a particular user. These constraints are applied and 46000 user access sequence are constructed from the log records over a week period.

- *Filtering Information using Data Mining*

Figure 2 indicates the IR process. This is achieved by applying data mining techniques, Content based filtering is done using association rule mining on the matrix. The results are interpreted n terms of conditional statements IF_THEN. Only single consequent rules are considered. if the same precondition occurs in more than one rule, the post-conditioned Web pages are ranked based on the confidence values of the rules. The content-based filtering rules are such that the preconditioned Web pages imply the post conditioned Web page based on the similarity in the keywords (i.e., textual content).

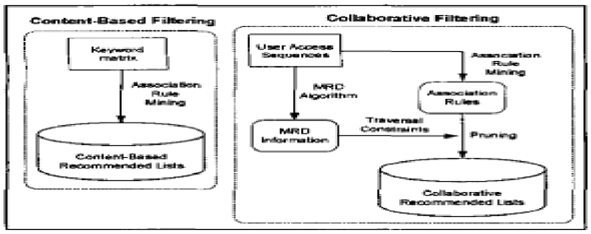


Fig 2 – Information Filtering using data mining technique

The collaborative filtering is also done in the Figure 2. This is achieved using the association rule mining on the server data of user sessions to generate a set of rules. The traversal constraint Minimum Reaching Distance is then constructed to capture the user behaviors on the web on a particular session. The experiments [5,6] shows that the performance in terms of precision and recall is increased when the MRD is used and also it outstands the Markov Model in mining the user access patterns.

- *DB design and UI development*

Database is designed using RDBMS . The data base consists of URLS, Keywords and recommended set of rules from content based filtering , user profiles and social filtering. Mysql is used as a query processing language as it provided multi user and multi threaded and robust SQL database management for recommender systems. The web interface is provided using the Apache web server and PHP. The database and the user interface are connected with a web server which takes request for user and sends back the needed pages. The interface is activated every time using a browser and via internet. The user profile resides on the server and user preferences are tracked . A logical path is attained between the user and their profile.

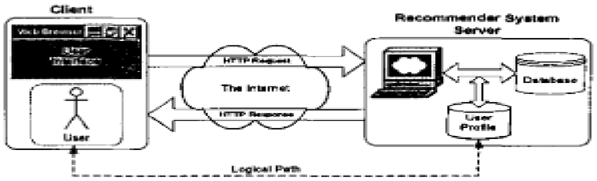
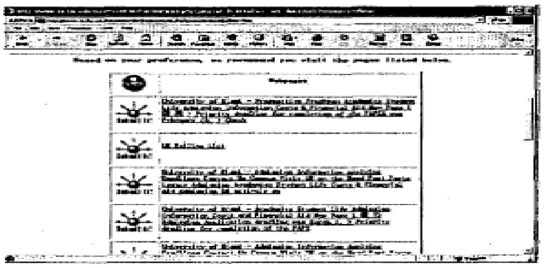


Fig 3:- Interaction of the User and the recommender system using HTTP

The recommended list of pages can be obtained from the content and social filtering. 10% of the recommended pages are from content based and the rest from social based. Figure 5 shows a snapshot of the recommender system interface by specifying 10% on the recommendation selection. The recommended list of Web pages which match the user's profile is shown on the bottom section of the page. To view the Web page, the user clicks on the URL name, and the actual Web page will be opened in a new window. The user clicks the button if the page is found to be useful in order to record his profile. By providing preferences, the system recommends web pages that are of user interest. Additional features are also added such as help and recent pages visited. The performance evaluation is made by making a group of users to participate in a survey.



Based on your preference, we recommend you visit the pages listed below.

Fig 4:- Snap Shot of the recommended system interface

Different group of users are involved in the evaluation process of the system by giving a score at a ten point scale. Most users agreed that by using the proposes system yield higher satisfaction than by the traditional query based matching algorithms. The average score was in the ratio 4:1 and also the ratio in terms of system ability to filtering the scale between content and social filtering was 4.3.

CONCLUSIONS

The lack of interaction between the user and the system, was the major back set in IR .The propose system enhances the interaction by analyzing the browsing patterns. The recommended pages are outcomes of individual users interest and preferences. A prototype of the system is proposed as a form of web navigational assistant. The recommender system helps in using the content and behavior. Thus additional information is presented to the users. The performance, response time are enhanced when compared with existing methods. In addition, the filtering scale-adjusting feature is also found to be useful to most of the users. The future work will be to develop a proto type for domain independent recommendation systems.

REFERENCES

[1]. M. Balabanovic and Y. Shoham, "Fab: Content-based, collaborative recommendation", Communications of the ACM, 40(3):6&72, 1997.

[2]. A data mining framework for building a Web-page recommender system C. Haruechaiyasak;M.-L. ShyuS.-C. Chen Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration, 2004. IRI 2004

[3]. C. Basu, H. Hirsh, and V. Cohen, "Recommendation as classification: Using social and content-based information in recommendation," Proc. of the Fifteenth National Conf. on Artificial Intelligence, pp. 7 14-720, 1998.

[4]. M. Chen, J. Han, and P. Yu, "Data mining: an overview from database perspective," IEEE Trans. on Knowledge and Data Engineering, 8(6):86&883, Dec. 1996.

[5]. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds., Advances in Knowledge Discovery and Data Mining, AAA1 Press, 1996.

[6]. C, Haruechaiyasak, M.-L. Shyu, and S.-C. Chen, "A Web-page recommender system via a data mining framework and the semantic Web concept," accepted for publication, International Journal of Computer Applications in Technology, Special Issue on Data Mining Applications, 2004.

[7]. C. Haruechaiyasak, "A data mining and semantic Web framework for building a Web-based recommender system," Ph.D. dissertation, University of Miami, June 2003.

[8]. P. Maes, "Agents that reduce work and information overload," Communications ofthe ACM, 37(7):3@40, 1994.

[9]. R. J. Mooney and L. Roy, "Content-based book recommending using learning for text categorization," Proc. of the Fifh ACM Con5 on Digital Libraries, pp. 195-204,2000.

[10]. B. M. Sarwarl G. Karypis, J. A. Konstan, and J. T. Riedl, "Item-based collaborative filtering recommendation algorithms," P roc. of the Tenth Int. WWw Cant, pp. 285-295,2001.

[11]. J. B. Schafer, J. A. Konstan, and J. Riedl, "Ecommerce recommendation application," Data Mining and Knowledge Discovery, 5(1/2): 115-153,2001.

[12]. R. Sedgewick, Algorithms in C, Addison-Wesley, 1990.

[13]. C. Shahabi, E Banaei-Kashani, Y.-S. Chen, and D. McLeod, "Yoda: An accurate and scalable Webbased recommendation system," Proc. ofthe Sixth Inr. Con8 on Cooperative Information Systems, September 2001.

[14]. U. Shardanand and P. Maes, "Social information filtering: Algorithms for automating word of mouth," Proc. of the Annual ACM SIGCHI on Human Factors in Computing Systems, pp. 210-217, 1995.

[15]. Y. Yang and J. P. Pedersen, "A comparative study on feature selection in text categorization," Proc. of the Fourteenth Inr. Conk on Machine Learning pp. 412- 420, 1997.

[16]. The Apache HTTP Server Project, http://httpd. apache. org [161 PHP: Hypertext Preprocessor. http://www.php.net [17] MySQL: Open Source Database, http://www.mysql.com.