ISSN No. 0976-5697

Volume 3, No. 1, Jan-Feb 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Web Mining Technique for Collaborative Web Surfing

Akhtar Ali Jalbani*, Gordhan Das Menghwar and Mukhtiar Memon
Information Technology Centre
Sindh Agriculture University Tandojam
Tandojam, Pakistan
[*akjalbani,gdas,mukhtiar.memon]@sau.edu.pk

Aneela Yasmin
Sindh Agriculture University Tandojam
Tandojam , Pakistan
aneelayasmin@sau.edu.pk

Abstract: Web mining is an application of a data mining technique which is used for knowledge discovery. In this paper, web mining techniques have been studied for the collaborative Web surfing, where more than one surfer are searching for the identical data from the world's largest database i.e. WWW. On the WWW the data is placed in an unstructured way. Therefore finding relevant information is always time consuming and a tiring job. We propose data mining technique using pattern matching method for collaborative web surfing.

Keywords: data mining; collaborative web surfing; web mining; HITS; knowledge discovery technique

I. INTRODUCTION

Data mining is a technique of extracting meaningful information from large and mostly unorganized data. It is the process of performing automated extraction and generating predictive information from huge data. The extraction of meaningful information from large data is otherwise known as knowledge discovery. There are varied views regarding the usage of term knowledge discovery for data mining. In this paper,we use data mining for web-based data as a unique process without bothering about various opinions of regarding the knowledge discovery [1, 2].

The data mining process uses different types of analysis tools for determination of relationship between data and for validation of predictive information. It is also integration of various techniques from multiple disciplines, for example, statistics, machine learning, pattern recognition, neural network, image process and data-management systems and so on [2].

Data mining technique is an evolving technology going through continuous modifications and enhancements. Security is an important issue associated with any data collection used for decision making. Organizing huge data, WWW data is a challenging task. The knowledge discovered by data mining tool is useful as long as it is interesting and understandable by the end user. Good data visualization makes easy and helpful for the user to interpret that data in a better way. In this paper, we propose data mining technique for collaborative web surfing, in which more than one surfer on the web is looking for the identical information. This approach is related to the World Wide Web hence web mining technique may be applied to understand the collaborative web surfing [3,4,5].

Furthermore, this paper is divided into six sections. Introduction is already presented in section I. Section II presents the concepts related to web mining. Pattern discovery for collaborative web surfing is discussed in Section III. Section IV presents a technique that how information can be retrieved using web mining for collaborative surfing. Application of HITS algorithm in the context of web surfing is discussed in Section V and Section VI presents summary and outlook.

II. WEB MINING

The measuring exact size of World Wide Web is extremely difficult. However, Google reported in 2001 [6] about the size of their own database, which was nearly 3 billion of the web documents. This database may be considered as the largest database available in the world. In this huge web data, various problems exist for example data is loosely coupled because no real structure has been adopted. Hence to identify knowledgeable or valuable data is much more difficult. To address these types of problem data mining approaches are quite useful for finding valuable data. The data mining has several applications. Our approach is related to web; therefore, we call it a web mining. In literature, [6] has classified web data into five essential factors, these are:

a. The data or content of the web page
b. The structure of the page linked outside includes HTML or XML code for the pages. This may be termed as Intra-page structure.
c. The structure of the page linked inside includes actual linkage structure between the web pages. This may be termed as inter-page structure.
d. How pages are accessed by the user; this may be termed as Usage data.
e. The information stored in cookies, for example, demographic, registration information and so on. This may be termed as user profiles.

Figure 1 shows the taxonomy of web mining activities described by [zai99]

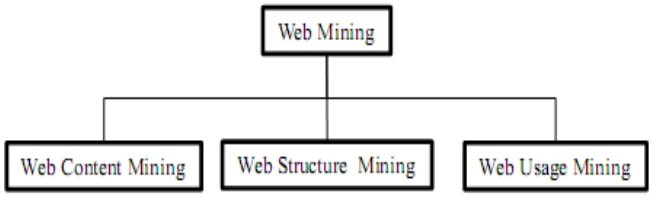

Fig.ure 1. Classification of Web Mining.

In this figure web mining is divided into three categories: Web content mining, web structure mining and web usage mining. Web content mining is further divided into the web page content mining and search result mining. The web usage mining is further divided into general access pattern tracking and customized usage tracking. There are many applications of web mining, for example, clustering may be applied to identify the web surfers visiting the same web page (Net Surf Utility) [7,8,9,10]. Generally it looks into the intra page structure where the HTML and XML code exists. The next step of web mining is to look into web access. This will go through the history to identify or track general access patterns. Normally, this technique is applied to the single user or specific user. In this example, the patterns should be identified using web mining techniques discussed earlier. The similar patterns can be identified and sent to the similar web site users to share their information. This also can be applied in a different way to cluster users into groups based on their similarity index.

The application of web mining may differ from domain to domain. To get more benefit from the application of web mining, we introduced this technique in collaborative surfing, where each surfer is looking for the similar information on the web. Hence, the users are visiting the similar web page to identify same information. To make more reliable and trustworthy information sharing process a web mining technique can play a vital role.

The effectiveness of web page setup depends upon structure of the web page, content and their ease of use. The effectiveness play major role to capture or attract more visitors on the web, this may also include various other parameters also, for example, user interface, graphics, response time and so on but in this paper we only discuss the techniques that can be adopted for mining for collaborative surfing, where more than one surfer is searching for particular information in a collaborative way.

A. Web Content Mining

The web content mining is something more than simple keyword searching on the search engines, or we can say it can be thought of an extending work performed by the basic search engines. On the internet, most of the search engines are based on keyword search techniques but web mining techniques improved this traditional way of searching on the search engines or crawling techniques such as indexing to store and query the information in a fast and reliable way. The data mining improves the efficiency, effectiveness and scalability of the search engines.

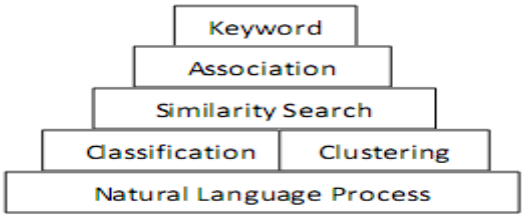

Figure 2. Text Mining Hierarchy.

In web mining techniques, the text mining is a simple and basic content mining technique. Figure 2 illustrates text mining functions bin hierarchy from top to bottom. Top are considered as simplest functions and bottom are become more complex functions. Much research is currently in progress to investigate the usage of natural language in the text mining. This will help to identify uncover hidden semantics, for example, questions and answer system. In web documents, retrieving data is still a problem because data is not structured in the traditional way as structured in the databases. It does not provide any schema or division of attributes. To extract data it needs more effort. HTML is semi structural language. Now this language will be replaced by XML. XML is the language which provides structured documents, and it will also be feasible to apply web mining techniques in an effective way.

B. Web Structure Mining

Web structure mining basically used to model the web organization in which web pages are classified according to similarity index. The main idea behind web structure mining is to improve the effectiveness of search engines and crawlers. However, page rank is another technique to improve effectiveness and efficiency of the web pages.

This mining technique can also be used for collaborative systems, where each visitor looks for the particular information by using keyword searches on the web. The web based collaborative software has been already developed, in which two similar keyword surfers can communicate with each other. Hence, the similarities are based on the keyword and surfers who are visiting the same URL.

C. Web Usage Mining

The web usages data or web log is the main entity of the web usage mining. The logs are maintained either from server or client perspective. Server side mining improves the design of the web sites. Web mining can also be used for the fetching of the web sites. In collaborative system, web mining plays both role clients as well as the server. Server maintains the logs, and client executes a sequence of clicks and information about those clicks are detected based on the client. The web mining helps to the collaborative systems for improving overall quality; effectiveness of the pages at the site can be easily evaluated [11].

The collaborative system for web usage mining is based on three general ideas:

a. Before providing the list of similar web site visitors, a reformatting of the web data is necessary before processing.

b. Pattern discovery based on exact keyword matching or parts of the keyword. This type of activity will help in identifying the hidden patterns in the web logs.
c. Pattern analysis, this process will help in interpretation of discovery results.

III. PATTERN DISCOVERY IN COLLABORATIVE WEB SURFING

Traversal patterns are the most common techniques in pattern discovery. The definition in terms of web page visiting may be summarized as set of web pages visited in one session is called traversal pattern. Similar pattern in web surfing can be identified by clustering similar traversal patterns. Several patterns can be identified, for example, duplicates page references or alternative of any page referenced in the same session. Patterns can be identified by using different combinations and knowledge of contiguous keyword or page references [12, 13].

Using Web mining technique web surfing information is processed from available data of each visitor. In this context, web mining discovers the similar web page visitors of the web surfer system. The discovery is based on pattern recognition, which extracts similar patterns from the available data. The pattern recognition method enables to surf and access data available in the netsurf utility more efficiently.

The pattern-matching technique for retrieving visitor based information from the server based to log files and applying this information for analyzing. Hence, this type of pattern matching is said to be web log mining, for example, visitors visited various types of web sites those are stored in the web log, in netsurf utility other pages that are embedded with the browser are chatting window. All the web pages are stored in the web log files.

In netsurf, web surfing tool, there are two types of web log mining files; these are: access logs, and agent logs. An access log file keep track of all documents that visitor has requested that include html files of the web page and their embedded graphic images, other associated files such as text files. An agent log file consists of records of the browser that was used to explore the web pages visited by the web surfer [7, 8, 9, and 10].

IV. INFORMATION RETRIEVEL SYSTEM FOR COLLABORATIVE SURFING

The information retrieval system for collaborative surfing consists of the web surfers and the URL of the web page document of the visiting web pages [6]. The information is stored in the set of documents. Hence, these may be represented as

$$U= \{U_1, U_2, U_3, \ldots\ldots, U_n\} \quad (1)$$

The input given as query of search purpose consist of key words, for example, query is represented by q and similarity between each URL present in the document set and the query is given calculated as:

$$\text{Similarity } (q, U_i) \quad (2)$$

The similarity function is set of membership function or the set, and it describes the similarity between the documents and the query is string given by the user. Now he issue is measuring the performance of information retrieval of the web surfing process that is done by two methods namely Precision and Recall.

Precision provides the required information,i.e. similarities between the URL visited by each user and the Recall measures or answer the question of does the matching process of the information retrieval process retrieve all documents matching the query.

V. HYPERLINK INDUCED TOPIC SEARCH (HITS)

The Hyperlink Induced Topic Search is a common method or an algorithm for knowledge discovery in the web. A HIT is a web searching method where the searching logic partially depends on hyperlinks to identify and locate the documents relating to the topic in the web. Simply, the HITS algorithm discovers the hubs and authorities of a community on the specific topic or query. The HITS algorithm accepts a set of web site reference as input. These input set called speed set and is returned by the web search engine. In HITS algorithm, the numbers of links between Web sites are measured as weights [6].

VI. SUMMARY AND OUTLOOK

The size and span of the WWW are huge and very wide. This it takes time to explore such a large volume of data. Web mining in collaborative web surfing may suffer due to the widely distributed; the communication breaks may occur. Interconnection of web pages is an issue, that may create difficulties in web mining of collaborative web surfing. Hidden information or web sources sometimes are difficult to explore. Hence web mining is challenging task for collaborative web surfing, in this paper, we have proposed an approach that may be applied using an integrative way using HITS algorithm to identify the mining activities in web surfing. Web mining can play the vital role in pattern identification of similar web page visitors. A machine learning approach can be considered as an outlook approach of collaborative web surfing using data mining techniques.

VII. REFERENCES

[1] R. Cooley, B. Mobasher, and J. Srivastave, "Web Mining: Information and Pattern Discovery on the World Wide Web" In Proceedings of the 9th IEEE International Conference on Tool with Artificial Intelligence, 1997

[2] D. Pierrakos, G. Paliouras, C. Papatheodorou, C.D Spyropoulos, "Web usage mining as a tool for personalization: a survey", User modelling and user adapted interaction journal, Vol.13, Issue 4, 2003, pp. 311–372

[3] B. Mobasher, H. Dai, T. Kuo, and M. Nakagawa, "Effective Personalization Based on Association Rule Discover from Web Usage Data" In Proceedings of WIDM 2001, Atlanta, GA, USA, pp. 9–15

[4] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic Personalization based on web usage Mining" Communications of the ACM, Vol. 43, No.8, 2000, pp. 142–151

[5] M. Eirinaki, M. Vazirgiannis, "Web Mining for Web Personalization", ACM Transactions on Internet Technology, Vol.3, No.1, February 2003

[6] G.H. Dunhm, "Data Mining: Introductory and Advanced Topics", Pearson Eduction Inc. 2003.

[7] A.A Jalbani, S. Abbasi, G.D Menghwar, and A. Yasmin, "Using Collaborative Enviornment for Web Surfing" Pak Journal of Agr. Engg. Vet. Sci. Vol27(1), 2011, pp-94-99.

[8] A.A Jalbani, S. Abbasi, G.D Menghwar, and A. Yasmin, "Towards an Approach for Web Surfing in Unison" In Proceedings of 4th Internation Conference on Development in e-Systems Engineering, UAE Dubai 2011.

[9] A.A Jalbani, G.D Menghwar, M. Memon, and A. Yasmin, "Usability of Collaborative Web Surfing Systems in e-Research" International Journal of Computer Science Issues, Vol9(1), 2012.

[10] A.A Jalbani, A. Yasmin, G.D Menghwar, and M. Memon, "Collaborative Web Surfing", Journal of Computing, Vol. 4(1), 2012.

[11] C. Lillian, T. I-Hsien ,K. Chris, W. Peter, K. Daniel "Combining ethnographic and clickstream data to identify user Web browsing strategies" Journal of Information Research, Vol. 11 No. 2, January 2006

[12] R. Kohavi, L. Mason, and Z. Zheng, "Lessons and Challenges from Mining Retail E-commerce Data" Machine Learning, Vol 57, 2004, pp. 83–113

[13] R. Baraglia, F. Silvestri, "Dynamic personalization of web sites without user intervention", In Communication of the ACM 50(2): 2007, pp-63-67

