IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. VIII (Mar-Apr. 2014), PP 134-137
www.iosrjournals.org

Hybridization of Web Content and Structure Mining (HWCSM) Technique by means of Content Based Ranking Algorithm

Supriya Byreddy[1], RajaniKanth Aluvalu[2]
[1,2]*Department of Computer Engineering, School of Engineering, RK University Rajkot*

Abstract: *Current Day scenario most of the applications moved on to web to enable any where computing. Mining approaches are proved as the best to extract knowledge. Organizations have to use various mining techniques to extract useful information. This information will help in day-to-day business functionality of organization. We can say web mining as the applications of the general data mining techniques to the Web. However, the internal properties of the Web force us to modify and extend the traditional techniques considerably. In this paper we are proposing an Approach to hybridize web content and web structure mining to improve the performance of web mining.*
Keywords: *Web mining, Content mining, Structure mining*

I. Introduction

Web mining is one of the application areas of data mining techniques to find out patterns from the Web [1, 2]. Data mining is defined as finding hidden information from the data stored in a database and therefore it has been called exploratory data analysis, data driven discovery, and deductive learning. There are of three types of mining: data mining, web mining, and text mining [18]. Web mining combines the two of the activated research areas i.e. Data Mining and World Wide Web. Therefore, Web mining can be defined as the application of data mining techniques in order to discover patterns from the Web data, comprising Web documents, hyperlinks between documents, and usage logs of the websites [10]. Web mining research overlaps substantially with other areas, including data mining, text mining, information retrieval, and Web retrieval [14]. Millions and millions of data are added up frequently to already present millions of data online. Because of this there is exponential growth of existing data online. Due to its fast and disordered growth, the World Wide Web has grown into a vast repository of online data with no appropriate organizational structure [6]. Predicting the users' preferences for improving the web use ability has become a challenging task.

There are three kinds of information that have to be dealt with when any user is accessing any web site [7].So the three types of information are based on content of data, structure of data and log data. Based on these three types information research area of web mining has been divided into web usage mining, web structure mining and web content mining [8, 9]. Technically Web usage mining is the process of extracting useful information from web server logs i.e. user's history [3]. Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web document text mining, resource discovery based on concepts indexing or agent based technology may also fall in this category. Web structure mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web [17]. Finally, web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in web access logs [11] .The main goal of Web usage mining is finding out what users are looking for on the Internet [4]. Some users might be searching over the internet specifically for textual data, while some other users may have the interest in multimedia data [5]. Here, we mainly concentrate on web structure and web content mining process.

Web structure mining is used to identify the relationship between the web pages linked by information and this is based on the web structure schema of the web pages. Structure mining allows search engine to identify and retrieve a search query directly to the linking web page from the web site the content is located. Structure mining is extensively used to extract previously unknown relationships between web pages. Based on the pattern of usage mined the navigation of web pages can be adapted to provide better user experience [13].Web content mining is the scanning and mining of text, pictures from web pages to determine the relevance of the content to the search query [12]. Web content mining is related but different from data mining and text mining. It is related to data mining because many data mining techniques can be applied in Web content mining. It is related to text mining because much of the web contents are texts. However, it is also quite different from data mining because Web data are mainly semi-structured and/or unstructured, while data mining deals primarily with structured data [15]. Web content mining is also different from text mining because of the semi-structure nature of the Web, while text mining focuses on unstructured texts. Web content mining thus requires creative applications of data mining and/or text mining techniques and also its own unique approaches [16].

1.1Statement of Problem

Web mining is the application of data mining methods used to discover relevant patterns from the Web data, including Web documents, hyperlink between documents, and usage logs of the websites. Web Content Mining, Web Structure Mining and Web Usage Mining are the three different categories of Web Mining. We focus on the web content and web structure mining techniques in web mining. In literature, different web content and web structure mining methods and algorithms were utilized in the mining process. Besides the other, the most recently developed Web Content Mining Using Clustering Technique mines the contents form the web documents more efficiently than the generalized association pattern algorithm. But, this technique has some drawbacks in the mining process i.e., first the relevant contents were grouped by clustering technique and then mining algorithms were applied on the clustered contents. This process takes more time and creates high complexity in the web content mining. The web structure mining process is normally performed by the rank algorithms namely, page rank, weighted page rank, and topic sensitive page rank. However, these page rank algorithms search-query results are independent of any particular search query. The drawback of page rank algorithms was reduced by the most recent algorithm called topic sensitive weighted page rank. In this page rank algorithm, the pages are ranked based on the user interaction to the web pages. This topic sensitive weighted page rank algorithm has considered only the user interaction factor in their page ranking process and so it does not produce the most relevant results. Moreover, combination of web content and structure mining related works were very less. If solutions are found for such drawbacks in the literary works, then the web content and structure mining process is improved with high accurate results. Hence, the lack of solution for such drawbacks has motivated to do the research work in this area.

II. Related Study

Kamlesh Patidar et al. [19] have discussed that the existing search engines do not provide content search from the collection of database, as no information about context was stored in database. The existing search engine mostly uses the agent based search and then the database based search. With the possibility of world wide access, current Web applications have an almost unlimited amount of potential users. With the collected information, changes and optimizations can be applied to the Web application to hit the user's requirements. The user who may be researcher, students, and even common person expect subject or context and need content accessibility precise and subject specific. As a field of research in data mining and knowledge discovery, today's digital library is a massive collection of various types and categories of documents. They have designed a prototype as a search engine and also proposed an algorithm for web content mining using the database approach and multilevel Data tracking for digital library.

Gauri Jain et al. [20] have proposed an Intelligent Model, which can help organizations to restructure their website so that website structure can be refined and it becomes more efficient and user friendly. In order to reach the main goal of web mining process, web mining algorithm has been applied. These can help to suggest the possible changes in the design of the website so that a common user feels much more comfortable in browsing the website. The data has been collected from a university website using a web crawler. Web mining algorithm has been applied on the data so that possible changes can be suggested.

Manikandan [21] has discussed that various association rule mining algorithms like Generalized pattern algorithm are being implemented to mine the web content but again due to the above setbacks the efficiency expected from the algorithm is not obtained. Since the dip in the efficiency of these algorithms is amounted to the nature of the textual web content, an algorithm which may deal with, if not all the anomalies at least the unclustered nature of the content may increase the efficiency drastically. They have made assumptions that the web content is static and there is at least one common pattern found in the given datasets.

LI Xiang-wei et al. [22] have proposed a web structure mining preprocessing algorithm based on the Rough Sets (RS). Firstly, to linear the huge web link graph, the Vast Forward Path (VFP) has been introduced and extracted from the user access record in web server logs. Secondly, to build the data analysis model, the Information System has been constructed using the VFP. Thirdly, the attribute reduction theory of RS has been used and the Information System has been reduced by eliminating a lot of abundant attributes. The experiments have shown that the proposed algorithm can get high efficiency and avoid the abundant web redundant data.

Shesh Narayan Mishra et al. [23] have proposed a topic sensitive weighted page rank algorithm based on web structure mining. The relevancy of the pages of a given topic was better determined, as compared to the existing Page Rank, Topic sensitive Page Rank and Weighted Page Rank algorithms. For ordinary keyword search queries, Topic Sensitive Weighted Page Rank scores satisfy the topic of the query.

III. Objectives

The main objective of this study is to provide efficient web content and structure mining method for obtaining accurate mining results from the web documents. The objectives of this study are highlighted as follows:

- To develop hybrid web content and structure mining technique for producing an accurate web mining result.
- To reduce processing time and to get more accurate results by making user feedback with user interaction process.
- To develop clustering based page rank algorithm for obtaining the order of web pages in the result list in order that the user may obtain the significant pages easily.

IV. Proposed Methodology

The main aim of this research is to provide a better hybridization technique by solving the drawbacks that currently exist in the literary works. Thus, we intended to propose a Hybrid Web Content and Structure Mining (HWCSM) technique to extract the accurate information from the web documents. In the proposed technique, initially the user will give a query request to the web documents, and then the query related pages and the corresponding links will be saved in the RI registry. Moreover, the queried user interaction and feedbacks will be collected and the user interaction time will be calculated. Subsequently, the RI registry pages with hyperlinks will be clustered and ranked by the clustering based page rank algorithm. The proposed clustering based page rank algorithm will utilize the calculated user interaction time and the collected user's feedback in their page rank process. Thus, our proposed technique will be more accurate than the existing page rank algorithm based WCM and WSM. Finally, the results will be analyzed to demonstrate the performance of the proposed HWCSM technique.

V. Possible Outcome

By using the proposed HWCSM technique, the user will easily get more accurate queried content from the relevant web pages. The performance of our proposed technique will be tested by using more number of user queries, and the results related to user queries will be compared with the conventional web content and structure mining techniques. Overall, by our proposed technique, the users can accurately mine the contents from the most relevant pages.

References

[1]. J. Vellingiri and S. ChenthurPandian, "A Novel Technique for Web Log mining with Better Data Cleaning and Transaction Identification", Journal of Computer Science,Vol. 7, No. 5, pp. 683-689, 2011

[2]. J.Indumathi and G.V. Uma, "Customized Privacy Preservation Using Unknowns to Stymie Unearthing Of Association Rules", Journal of Computer Science, Vol. 3, No. 11, pp. 874-881, 2007

[3]. Tzung-Pei Hong, Ming-Jer Chiang, Shyue-Liang Wang, "Mining Fuzzy Weighted Browsing Patterns from TimeDuration and with Linguistic Thresholds", American Journal of Applied Sciences, Vol. 5, No. 12, pp. 1611, 2008

[4]. H. Parveen Sultana, M. Pounambal and Dr. P. Venkata Krishna, "A Fast Handover Scheme for Multicasting in IPv6 basedMobile Ad hoc Networks", Journal of Computer Science, Vol. 7, No. 1, pp. 90-94, 2011

[5]. Chandramathi, N. Chandra Sekhar, M.G. Adarsh. Girish Chandra Haomom and Thirumalaivasan, "A Novel Video Surveillance System Based on Multimedia Messaging Service", Journal of Computer Science, Vol. 1, No. 2, pp. 244-248, 2005

[6]. Muhammad Shoaib and Abad Ali Shah, "A Methodology to Segment the Text for Index Terms", American Journal of Applied Sciences, Vol. 2, No. 9, pp. 1309-1314, 2005

[7]. Haider Ramadhan, Muna Hatem, Zuhoor A1-Khanjri and Swamy Kutti, "A Classification of Techniques for Web Usage Analysis", Journal of Computer Science, Vol. 1, No. 3, pp. 413-418, 2005

[8]. Vijayakumar Mohanraj and Muthaial Chandrasekaran, "An Ontology Based Approach to Implement the OnlineRecommendation System", Journal of Computer Science, Vol. 7, No. 4, pp. 573-581, 2011

[9]. YahyaAlMurtadha, Md. Nasir Bin Sulaiman, Norwati Mustapha and NurIzuraUdzir, "IPACT: Improved Web Page Recommendation System Using ProfileAggregation Based On Clustering of Transactions", American Journal of Applied Sciences, Vol. 8, No. 3, pp. 277-283, 2011

[10]. Surbhi Anand and Rinkle Rani Aggarwal, "Data Mining Types and Techniques: A Survey", International Journal of Research in IT & Management, Vol. 2, No. 2, pp. 458-471, 2012

[11]. Kavita D.Satokar Gawali, "Web Personalization Using Web Mining", International Journal of Engineering Science and Technology, Vol. 2, No. 3, pp. 307-311, 2010

[12]. Bin He, Kevin chen-chuan chang, "Automatic complex schema matching across web query interfaces: A correlation mining approach", ACM Transactions on Databases Systems; Vol. 31, No.1, pp. 1-45, March 2006

[13]. Subramanian and Abdul Rauf., "Structure Mining for Web Link Recommender System", International Journal of Computer Science and Technology, Vol. 2, No. 4, 2011

[14]. Hsinchun Chen and Michael Chau, "Web Mining: Machine learning for Web Applications", Annual Review of Information Science and Technology, Vol. 38, pp. 289-329, 2004

[15]. Venkata Ramana Adari, Diwakar and Suresh Varma, "Face Location - A Novel Approach to Post the User global Location", Control Theory and Informatics, Vol. 1, No. 1, pp. 34-38, 2011

[16]. Udayasri.B, Sushmitha and Padmavathi, "A Lime Light on the Emerging Trends of Web Mining", f International Journal of Computer Science & Informatics, Vol. 2, No. 1, 2, pp. 65-70, 2012

[17]. Ashish Sharma and Niket Bhargava, "An Approach to Enhance Web Service Resource Framework using the Improved PLWAP Algorithm for Large Scale Hybrid Data in Distributed Environment", International Journal of Computer Technology and Electronics Engineering, Vol. 1, No. 2, pp. 90-95

[18]. Sankar K. Pal, Varun Talwar and Pabitra Mitra, "Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions", IEEE Transactions on Neural Networks, Vol. 13, No. 5, pp. 1163-1177, September 2002

[19]. Kamlesh Patidar, Preetesh Purohit and Kapil Sharma, "Web Content Mining Using Database Approach and Multilevel Data Tracking Methodology for Digital Library", International Journal of Computer Sci ence and Technology, Vol. 2, No. 1, pp. 194-198, March 2011
[20]. Gauri Jain and Varun Kumar, "An Intelligent Model for Redesigning Websites using Web Mining Techniques", International Journal of Computer Applications, Vol. 25, No. 5, pp. 14-18, July 2011
[21]. Manikandan, "Improving Efficiency Of Textual Static Web Content Mining Using Clustering Techniques", Journal of Theoretical and Applied Information Technology, Vol. 33, No. 2, pp. 193-196, 2011
[22]. LI Xiang-wei, Zheng Gang and Kang Yu-xue, "A Rough Sets Based Data Preprocessing Algorithm for Web Structure Mining", International Technology Journal, pp. 1-4, 2012
[23]. Shesh Narayan Mishra,Alka Jaiswal and Asha Ambhaikar, "An Effective Algorithm for Web Mining Based on Topic Sensitive Link Analysis", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, No. 4, pp. 278-282, 2012