IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

Comparative Study and Analysis on the Techniques of Web Mining

Dipika Sahu[1], Yamini Chouhan[2]

Department of Computer Science and Engineering, Faculty of Engineering and Technology, Shri Shankaracharya Technical Campus, Junwani, Bhilai, District-Durg, Chhattisgarh, India[1, 2]

Abstract: Now a day's most of the people rely on internet. At the same time internet has many information. It should give related information for each user query. Web mining is used to extract information based on the user query from the large collection of data accessible in web. It is concerned mainly with its content, structure and usage. Web usage mining is extracting information based on user log, frequently accessed paths. Web content mining is used to fetch information from the web files. Web structure mining usually use graph theory to extract the web site structure through which they give better search results for the user. This paper also reports the summary of various techniques of web mining approached from the following angle like Feature Extraction, Transformation and Representation and Data Mining Techniques in many application domains.

Keywords: Web Mining, Web Usage Mining, Web Structure Mining, Web Content Mining, Graph theory.

I. INTRODUCTION

The web is extremely enormous, diverse, flexible, and dynamic. The World Wide Web continues to develop both in huge volume of traffic and the size and complexity of Web sites. With the increasing growth of information accessible in net, it is difficult to identify the relevant information present in the web. Meanwhile much information is unstructured. It is essential to use automated tool for obtaining the necessary information from the huge collection of information.

Web Mining is used to extract information from the raw unstructured data. The emerging field of web mining objective at finding and extracting relevant information that is hidden in Web related data, in particular in text files published on the Web. Web mining is performed in three ways they are 1) web usage mining 2) web content mining 3) web structure mining. Web usage mining gives the support for the web site design, giving personalization server and the other business making decisions, etc. Web content mining is the process of extracting knowledge from the content of files or their descriptions. Web document text mining, resource discovery based on concepts indexing or agent; based technology might also fall in this category. Web structure mining is the process of inferring knowledge from the World Wide Web organization and the link between references and referents in the Web. Finally, web usage mining, also known as the Web Log Mining, is the process of extracting interesting patterns in web access logs. In order to superior serve for the user, web mining apply the data mining, the artificial intelligence and the chart technologies and so on to web data and traces user visiting characteristic, and then extracts the users' using pattern. It has rapidly become one of the most significant areas in Computer and Information Sciences because of its direct applications in the e-commerce, CRM, Web analytic, information retrieval and filtering, and Web information systems.

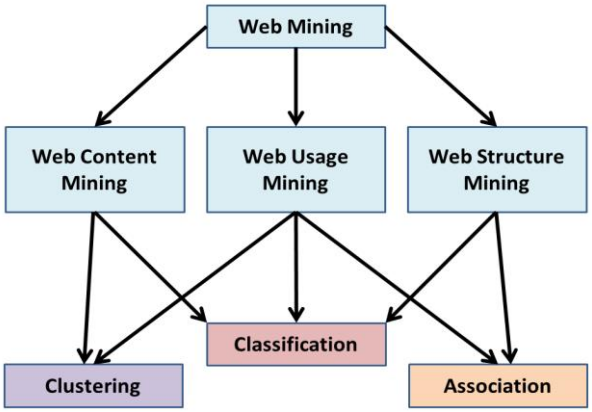

Fig 1 – Taxonomy of Web Mining

Web mining is the utilize of data mining techniques to automatically extract data from the web. Web mining has several sub tasks they are:-

1) Resource finding
2) Information Selection and pre-processing
3) Generalization
4) Analysis

Resources finding is the task of retrieving intended Web files. Information Selection and pre-processing is automatically selecting and pre-processing precise information from retrieved web resources. Generalization is automatically selecting and pre-processing precise information from retrieved wed resources. Analysis is validation and interpretation of mined patterns.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

Web usage mining is utilized to extract information based on the user log. Web Usage mining is the process of applying data mining techniques to the finding of usage patterns from Web data, targeted towards various applications. Discovery of meaningful patterns from the data are generated by client-server transactions on one or more web servers.

Typical Sources of Data are Automatically produced data stored in server access logs, referrer logs, agent logs, and client-side cookies, E-commerce and the product-oriented user events (e.g. shopping cart changes, ad or product click-through, etc), Users profile and/or users rating and Meta-data, page attributes page content, site structure.

II. PROBLEM IDENTIFICATION

The web is extremely dynamic; lots of pages are added, updated and removed everyday and it handles huge set of information hence there is an arrival of several number of problems or issues. Normally, web data is high dimensional, limited query interfaces, keyword oriented search and restricted customization to individual users. Due to this, it is very difficult to find the relevant information from the web which may generate new issues. Web mining techniques are classification, clustering and association rule which is utilized to understand the customer behavior, evaluate a particular web-site by utilizing traditional data mining parameters. Web mining procedure is divided into four steps; they are resource finding, data selection and pre-processing, generalizations and analysis [11] [8]. Web measurement or web analytics are one of the significant challenges in the web mining. The measurement factor are hits, page views, visits or user sessions and find the unique visitor regularly utilized to measure the user impact of various proposed changes. Large institution and organization archive usages data from the web sites [10]. The major problem is that, detecting and/or preventing fraud activity. The web usage mining algorithms are more efficient and precise. But there is a challenge that has to be taken into the consideration. Web cleaning is the most significant process but data cleaning becomes difficult when it comes to heterogeneous data [10]. Maintaining accuracy in classifying the data wants to be concentrated. Although various classification techniques exist the quality of clustering is the still a questions to be answered.

The Major issues in the process of web mining are:-

- Web data sets can be very big; it takes ten to hundreds of terabytes to store on database.
- It can't mine on a single server so it wants large number of server.
- Proper organization of the hardware and the software to mine multi-terabyte data sets.
- Limited customization, restricted coverage, and limited query interface to the individual users.
- Automated data cleaning.
- Over fitting and under the fitting of data.
- Over sampling of the data.
- Scaling up for high dimensional data.
- Mining series and time series data.
- Difficulty in finding related information.
- Extracting new knowledge from the web.
- Data / Information Extraction concentrate on extraction of the structured data from web pages such as products and the search result.
- Web information integration and schema matching. The web contains huge amount of data, each web site accept similar information in a diverse way. Similar data discovery is an significant problem with lots of the realistic applications.
- Opinion extraction from the online sources i.e. customer makes sure of products, forums, blogs and chat room. Mining opinions are of large consequence for marketing intelligence and the product benchmarking.
- Automatically segmenting web pages and identifying noise is an interesting trouble in Web application. It could not have advertisements, navigation link and the copyrights notices. Hence, extracting the main contents of web page is important problem in web application.

III. METHODOLOGY

A) Data collection: Data collection is the initial step of web usage mining, the data authenticity and the integrality will directly affect to the following works smoothly carrying on and the final recommendation of characteristics service's quality. Therefore it must utilize scientific, reasonable and advanced technology to gather many data. At present, towards web usage mining technologies, the main data origin has three kinds: server data, client data and the middle data.

B) Data preprocessing: Some database is In - sufficient, inconsistent and including noise. The data pretreatment is to carry on a unifications transformation to those databases. The result is that the database will to become integrated and consistent.

C) Knowledge Discovery: Use statistical method to carry on the analysis and mine the pretreated data. We might discover the user or the user community's interests then construct interest model. At present the generally used machine learning methods mainly have clustering, classifying, the relation discovery and order model discovery. Each method has its own excellences and shortcomings, but the quite effective method mainly is classifying and clustering at the presents.

D) Pattern analysis: Challenges of Pattern Analysis are to filter uninteresting information and to visualized and interpret the interesting pattern to the users. Initial delete the less significance rules or models from the interested model storehouse; Next utilize technology of OLAP and

so on to carry on the comprehensive mining and analysis; Once more, let discovered information or knowledge be visible; Finally, provide the characteristic service to the electronic commerce website.

E) Focused Crawling: A focused web crawler takes a set of well-selected web pages exemplifying the user interest. The focused crawler start from the given page and recursively explores the linked web pages. While the crawlers performs a breadth-first explore of the complete web, a focused crawler explores only a small portion of the web using a best-first search guided by the user interests. crawling for retrieving multimedia content in the web, instead of plain HTML documents.

F) Clustering Web Objects: Focused crawling retrieves large numbers of relevant data. In order to offer fast and more precise access to the query results, clustering is an established method to group the retrieved statistic to achieve superior understanding. If the query result is websites or combined objects like images and their text descriptions, algorithm are wanted to handle these combined data types to find meaningful clustering.

G) Wrapper Induction: A wrapper is a piece of software that allows a semi structured Web source to be queried as if it were a databases Given a sets of manually labeled page, a machine learning technique is applied to learn extraction rules or patterns.

H) Automatic Data Extraction: Given a set of positive page, produce extraction patterns. Given only a single page with multiples data record, create extraction patterns.

IV. CONCLUSION

In this paper we survey the research area of Web mining, focusing on the category of Web structures mining. We had introduced Web mining. Later in the paper when we had discussed Web structure mining, and introduces Link mining, as well as block-level links mining issues. We had also reviewed two popular algorithms to have a thought about their application and effectiveness. Since this is a huge area, and there a lot of work to do, we expect this paper could be a useful starting point for identifying opportunities for further research.

REFERENCES

[1] Ms. Dipa Dixit, Fr.CRIT, Vashi, M Kiruthika," PREPROCESSING OF WEB LOGS", (IJCSE) International Journal on Computer Science And Engineering, Vol. 02, No. 07, 2010, 2447-2452.
[2] Dr. Sohail Asghar, Dr. Nayyer Masood," Web Usage Mining: A Survey On Preprocessing Of Web Log File Tasawar Hussain", 978-1-4244-8003-6/10@2010.
[3] Theint Theint Aye "Web Log Cleaning For Mining Of Web Usage Patterns".
[4] S. K. Pani, et.al L "Web Usage Mining: A Survey On Pattern Extraction From Web Logs", International Journal Of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1, 2011.
[5] Chidansh Amitkumar Bhatt · Mohan S. Kankanhalli, "Multimedia Data Mining: State Of The Art And Challenges" Published Online: 16 November 2010© Springer Science+Business Media, LLC 2010.
[6] Margaret H. Dunham, Yongqiao Xiao Le Gruenwald, Zahid Hossain," A SURVEY OF ASSOCIATION RULES Web Usage Mining".
[7] Brijendra Singh1, Hemant Kumar Singh2,"WEB DATA MINING RESEARCH: A SURVEY", 978-1-4244-5967-4/10/$26.00 ©2010 IEEE.
[8] Rajni Pamnani, Pramila Chawan 1 Qingtian Han, Xiaoyan Gao, "Web Usage Mining: A Research Area In Web Mining".
[9] Wenguo Wu, "Study On Web Mining Algorithm Based On Usage Mining", Computer- Aided Industrial Design And Conceptual Design, 2008. CAID/CD 2008. 9th International Conference On 22-25 Nov.2008.
[10] R. Kosala, H. Blockeel. "Web Mining Research: A Survey," In SIGKDD Explorations, ACM Press, 2(1): 2000, Pp.1-15.
[11] http://www.kdnuggets.com
[12] http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.91.1602
[13] J Vellingiri, S.Chenthur Pandian, "A Survey on Web Usage Mining", Global Journal of Computer Science and Technology .Volume 11 Issue 4 Version 1.0 March 2011.
[14] Chen L, Mao X,Wei P, Xue Y, Ishizuka M (2012) Mandarin emotion recognition combining acoustic and emotional point information. Appl Intell 37(4):602–612.
[15] Shang F, Jiao LC, Shi J, Wang F, Gong M (2012) Fast affinity propagation clustering: a multilevel approach. Pattern Recognition 45(1):474–486.
[16] J. Shao, X. He, C. Bohm, Q. Yang, C. Plant, "Synchronization-Inspired Partitioning and Hierarchical Clustering," IEEE Transactions on Knowledge and Data Engineering, 2012.
[17] Ta¸sdemir K (2012) Vector quantization based approximate spectral clustering of large datasets. Pattern Recognition 45(8):3034–3044.
[18] Abdelhakim Herrouz, Chabane Khentout, Mahieddine Djoudi,''Overview of Web Content Mining Tools", The International Journal of Engineering And Science (IJES), Volume 2, Issue 6, 2013.