International Journal of Trend in Research and Development, Volume 4(6), ISSN: 2394-9333
www.ijtrd.com

A Study on Applications, Approaches and Issues of Web Content Mining

N. R. Satish
Assistant Professor, Department of Computer Science, University P.G. College O.U., Secunderabad, Telangana, India

***Abstract*—**Internet has become most popular resource of information in this technological era, as the expansion of web increases, humongous amount data and different kinds of information are storing online at the fast pace. With this past faced spreading of web data, it became more difficult to extract useful information from web. Web mining is a sub process of data mining which operates on web data. Web content mining is a method of web data mining or web mining. Web content mining primarily focuses on congregating, classifying, orchestrating of web data and furnishing the enhanced information from online entreated by user. This paper presents significant survey and analysis of web content mining methods and applications.

Keywords*—Web Mining; mining; Web Content Mining; Information; Knowledge; Web User;*

I. INTRODUCTION

Internet has been going through volcanic expansion in the span of past few years. Humongous sources of data available online, with large set of structured data and unstructured data added to web every day it is difficult to extract desired information from web. Use of efficient data extraction methods for retrieving data online is a imperatively significant need for users [7]. Data mining is a handy approach in this kind of situations. Predominantly data mining is referred as an activity which uncovers user required sequences or intelligence from divergent data sources like text, images, logs, online, and or database etc. discovered knowledge must be well founded, easy to understand, and possibly the right information required by user.

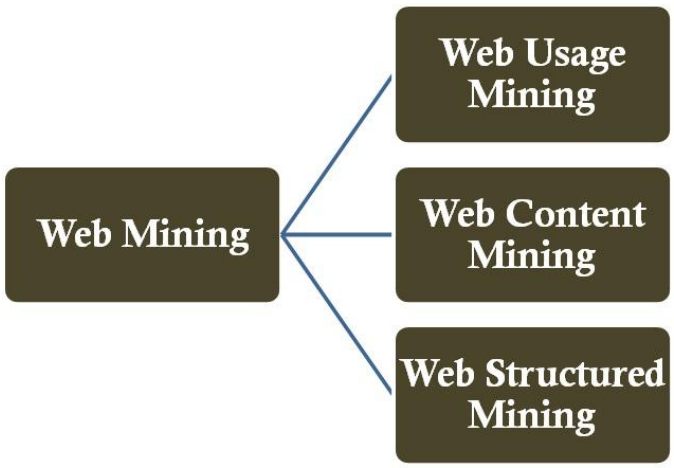

Fig. 1 Classification of Web Mining Techniques

Web mining is an action of data mining for finding useful knowledge from online data. This data may be web pages which are hyperlinked by other web pages, various inline documents, web logs, online videos and so forth. At first web mining was introduced by Etizoni [8] in the year 1996. At the outset he made presumption about web mining and its approaches in such a way that the information on web is organized adequately and meaningfully. He further stated that web mining is method which brings out much needed information from the internet sources. There are two ways to elucidate Web mining: one is process-centric view approach and the other is data-centric view approach. According to former approach web mining is a series of jobs and the later one defines it as method of mining processes operated on web data. As per the analysis, web mining can be split up into three types one is Web usage mining, second one is Web content mining and the last Web structure mining [1][21]. The following figure (Fig. 1) shows the classification web mining processes.

A. Web Usage Mining

Web Usage Mining vigorously focuses on methods which forecasts user exploration knowledge when they are online. Its primary focus is to track the user behavior based on generalized and personalized web usage. In generalized tracking knowledge is discovered according to web page history of user and personalized tracking, knowledge is extracted from particular user web interaction. Principally the web usage mining data sources are client logs, browser logs, server logs and proxy server logs [22] [23]. The main task of web usage mining is data preprocessing which compromises of data cleaning, identifying and building user sessions, getting web page information, and formatting the data. After the completion of data preprocessing it applies data mining algorithms to predict the user behavior.

B. Web Content Mining

The central point of web usage mining is to extracting knowledge from web page contents. The knowledge that can be extracted from web page contents is like product details on ecommerce site, social media postings and so on. This data can be used for many purposes such as tracking online terrorist activities, customer reviews about a product, and furthermore which is not possible with standard data mining approaches. This method incorporates integrates browsing records of users which constitutes of web page content, multimedia information like audio and video content. At end of this process it produces a list of organized or unorganized reports. Following sections in this paper presents a view of various approaches, applications and methods of web content mining.

C. Web Structred Mining

This method finds hidden relationships over online. It catalogues web pages to produces needful knowledge like resemblance and interrelation between pages. This method reviews hyperlink structure of web. Web Structure Mining interpreted as finding the structured knowledge from web link structure [22] [23]. Focusing on is the significant challenge for web link structure is main challenge of web structure mining. The primary purpose of web structure mining is to analyze the web pages and arrange them in a structured manner.

II. WEB CONTENT MINING APROACHES

First, Web content mining is methods that retrieve data from internet and process it produce well formed structures and arrange the data in a way that finding of required knowledge from web servers would be done at rapid pace. It includes discovering structured data from web sources, identifying and integrating similar data which is extracted. Since there is no link structure in a relational table of usual data mining approaches, this is not possible with those methods. There various approaches to this in accordance with different aspects such as structured data like html content, documents and databases, unstructured data like text. Mostly used web content mining is unstructured content mining, structured mining, Semi-structured content mining, and multimedia content mining. Figure fig. 2 describes the classification of web content mining.

A. Unstructured content Mining

The Majority of data on the web is in the form of unstructured text data. To process this data web content mining needs data mining and text mining techniques [2]. The result of unstructured data mining produces unknown knowledge [3]. Most commonly used text mining approaches are tracking the topic, summarization, categorization, clustering information extraction, and information visualization.

1) Information Extraction: To Important phrases and relationships are recognized over the textual data in this process. This process called as pattern matching and operates on for predefined sequences in the text. This method can also applied to vast amount of textual data.

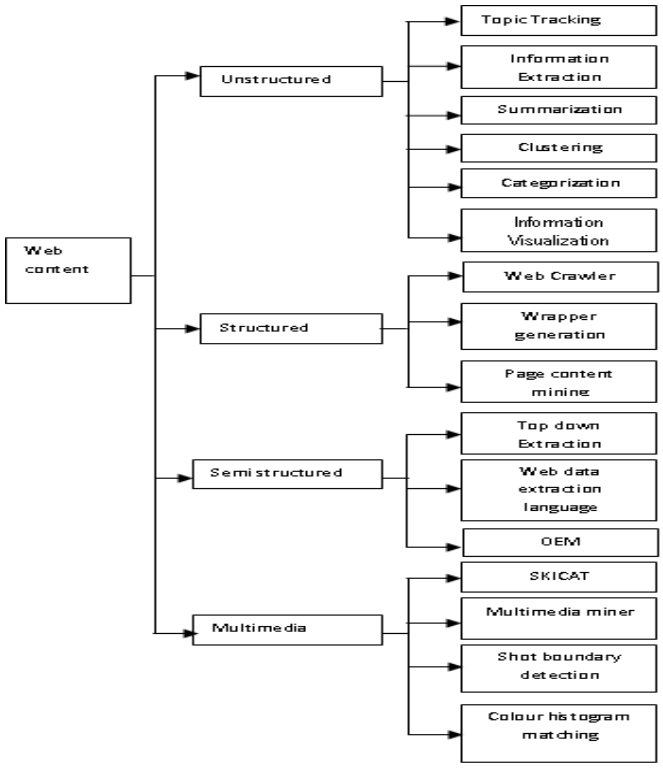

Fig. 2 Classification of Web Content Mining Approaches

2) Topic Tracking: Topic tracking approach can be adopted in text mining process. Its vital objective is to discovering and tracking of incidents existed in several news sources for example radio and TV airings. To make it easy for user this process tracks all the information consecutively and produces understandable knowledge.

3) Summarization: Extensive documents are summarized into a structure in this process. This method summarizes textual data and allows user to read the first paragraph of the document.

4) Categorization: This method is also referred as classification of text, or topic spotting. Categorization process sorts set of documents into various categories from a preordained set.

5) Clustering: Clustering forms groups of likewise documents. These groups contain documents that are logically same information and they are physically stored together. Disk access events are reduced to enhance the efficiency of database.

6) Information visualization: To strengthen the user intelligence, visual representations of abstract data are studied in this approach. Vast amount of textual content is visually analyzed by user in information visualization mappings. This process is attentive when user requires to see very large documents. With the creation of sub-maps, zooming and scaling user can interact with the document.

B. Structured Content Mining

Extraction of structured data is easier compared to unstructured data. Host pages are referred as structured data on the web. This process concludes three techniques one is web crawler, second one is wrapper generation, and the last one is page content mining technique. The program that extracts the structured data is called wrapper. In general data records those are retrieved from beneath database and visual content using predefined structure in a web page termed as structured data [9]. This process helps user to organize the data from various kinds' of sources.

1) Web Crawler: Web crawlers also mentioned as web spiders are actually intelligent software programs which searches over the web for desired knowledge. Search engine can achieve fast search results using web crawler, this is done by duplicating the all browsed web pages data for further actions and by indexing pages which are downloaded. It also updates the web content and their indexes automatically. For information extraction web crawlers make use of methods like breadth-first search, particle swarm optimization, genetic algorithm and/or other soft computing methods. And they also navigate through hyper links of web structure to extract knowledge.

2) Wrapper Generation: Wrapers serves meta information such a source domain, index links, statistics . In this process, based on the capability of source (queries that will be answered and result types) knowledge is provided. Wrapper extract content from specific data sources to translate it into link structured format. There are two way to do wrapper generation one is induction and the second one is automatic data extraction. Supervised learning used in induction for learning information extraction from trained programs provides by user. lots of the web data is in generic form so automatic data extraction is possible.

3) Page Content Mining: This process classifies the web pages. It is a structured approach of web data mining. This method operated by page ranking given by standard search engines.

C. Semi Structured Content Mining

Semi-structured data progressed from numbered relational tables and from strings. In semi structure mining process, real world complex objects are represented as natural representation without dislocating application writers. Evolution representations of semi-structured data are alternative forms of the OEM (Object Exchange Model). The data is in the form of compound (ex: labeled edges) or granular

(number or string) objects. Semi structured content mining has sub methods such as OEM extraction, Top down extraction, and table extraction. In OEM method object structure is self defined, in Top down extraction method complex objects are extracted from web sources and process them into understandable structures before the extraction of atomic objects and in the tables are extracted from web pages and web documents.

D. Multimedia Content Mining

Multimedia content mining usually a process of extracting knowledge from multimedia content such as audio/videodata and sound data. Multimedia content has various techniques such as SKICAT (Sky Image Cataloging and Imaging Tool) - In which large raw digital data is analyzed to produce is scientific knowledge, second one is Color Histogram Matching-is an image processing technique for matching the relationship between images based generated histograms of processed images. Multimedia Miner and Shot Boundary Detection techniques used to process the video content, later detects transitions between two frames where as former analyzes multimedia content. Multimedia miner processes multimedia data to extract image and video content, stores extracted images and video content into a database so that while performing discovery process user query can get desired results [19].

III. APPLICATIONS AND ISSUES OF WEB CONTENT MINING

Web content mining is used for grouping, classifying, arranging and producing the most useful information available on the internet. Also determines its relevance to the user query. Useful for the online marketing by enhanced exploration of information on the web. User can get highly refined information using Web content mining. Analyzes productiveness of websites, and tracks user online behavior, and helps digital marketing through predicting intelligence of product price, popularity etc. Customer reviews can be analyzed based on theses reviews product details can be extracted (about product and its performance). Users who have same interests are grouped and based on analyzing the content that they are posted on social media sites. Online content must be optimized by web mining since vast amount of content is added to the World Wide Web each every day. Cloud which handle with numerous files, images, videos and other large content must be optimized. The problem is with online data sets which require massive storage space to store in a database. Mining of single server is not useful so it requires more number of servers to process and to extract useful knowledge. Special software and hardware are obligatory to mine terabytes of data sets. There may be chance of deleting new data from the web with automated cleaning process. Limited customization, constrained scope, and restricted inquiry interface to individual clients. Some user may get more information than he required or some time less and sometimes there may be chance of extent refining of data than prescribed. Sometimes it is difficult to find significant knowledge on because content on web changes dynamically.

CONCLUSION

Internet has become most meaningful source of information nowadays. The information on the internet is in the form of web pages. Web pages contain almost more than fifty percent of noise data. Web content mining is process of extracting desired information from web by separating relevant content from non-relevant content. This paper explores various approaches of web content mining, applications and issues of web content mining. This process scans web page information such as HTML content, image, audio, and video content, and textual data. Outcome of this process supplied to search engines to produce more relevant knowledge. Since web contains dynamic data it is a complex process to mine the data that is added each every day. In Future, web content mining improves its approaches to enhance usability by predicting the requirements of user.

References

[1] V. David Martin, Dr. T. N. Ravi, “A Literature Survey on Web Content Mining”, IJRITCC, Volume 4, Issue 10, October 2016.
[2] Anurag kumar, Ravi Kumar Singh,” A Study on Web Content Mining”, IJECS, Volume 6, Issue 1, Jan 2017.
[3] Ms.S.Valarmathi, Mr.P.Purusothaman, “A Survey on Web Content Mining Techniques and Tools”, IJISET, Volume 1, Issue 6, August 2014.
[4] Anu, “Web Mining Evolution & Comparative Study with Data Mining”, IJRITCC, Volume 5, Issue 5, May 2017.
[5] D. Saravanan, N. Sugavaneswaran, “A Survey on Web Content Extraction Techniques”, INJRI, Vol 3, Issue 2, August 2016.
[6] Ananthi.J, “A Survey Web Content Mining Methods and Applications for Information Extraction from Online Shopping Sites”, IJCSIT, Vol. 5, 2014.
[7] Shohreh Ajoudanian, and Mohammad Davarpanah Jazi, “Deep Web Content Mining”, IJCEACIE, Vol 3, 2009.
[8] Surbhi Sharma, Dinesh Soni, Dr. Arvind K Sharma, “Explorative Study of Web Data Mining Techniques and Tools: A Review”, IJCST, Vol. 8, Issue 1, Jan - March 2017.
[9] Narendra Parmar, Dr. Vineet Richhariya, Jay Prakash Maurya, “An Exploratory Review of Web Content Mining Techniques and Methods”, IJARCCE, Vol. 5, Issue 5, May 2016.
[10] Abdelhakim Herrouz, Chabane Khentout, Mahieddine Djoudi, “Overview of Web Content Mining Tools”, IJES, Volume 2, Issue 6, 2013.
[11] Arvind Kumar Sharma, P.C. Gupta, “Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining”, IJARCET, Volume 1, Issue 8, October 2012.
[12] T. Shanmugapriya, P. Kiruthika, “Survey on Web Content Mining and Its Tools”, IJSER, Volume 2, Issue 8, August 2014.
[13] Zahra Hojati, Rozita Jamili Oskouei, “A Comprehensive Comparison between Web Content Mining Tools: Usages, Capabilities and Limitations”.
[14] R.Malarvizhi, K.Saraswathi, “Web Content Mining Techniques Tools & Algorithms – A Comprehensive Study”, IJCTT, volume 4, Issue 8, August 2013.
[15] CLAUDIA ELENA DINUCĂ, DUMITRU CIOBANU, “WEB CONTENT MINING”, Annals of the University of Petroşani, Economics, 12(1), 2012.
[16] Karan Sukhija, “Web Content Mining equipped Natural Language Processing for handling web data”, IJCAT, Volume 4, Issue 3.
[17] Govind Murari Upadhyay, Kanika Dhingra, “Web Content Mining: Its Techniques and Uses”, ijarcsse, Volume 3, Issue 11, November 2013.
[18] Deepti Sharda, Sonal Chawla, “WEB CONTENT MINING TECHNIQUES : A STUDY”, IJIRTS.
[19] Faustina Johnson, Santosh Kumar Gupta, “Web Content Mining Techniques: A Survey”, ijca, Volume 47, No.11, June 2012.
[20] R.Malarvizhi, K.Saraswathi, “Web Content Mining Techniques Tools & Algorithms – A Comprehensive Study”, IJCTT, Volume 4, Issue 8, August 2013.
[21] Feiran Huang, Jia Li, Jiaheng Lu, Tok Wang Ling, Zhaoan Dong, "PandaSearch: A Fine Grained Academic Search Engine for Research Documents", 2015 IEEE 31st International Conference on Data Engineering, 1408 – 1411.
[22] Kamika Chaudhary, Santosh Kumar Gupta, “Web Usage Mining Tools & Techniques: A Survey”, IJSER, Volume 4, Issue 6, June 2013.
[23] P. Britos, D. Martinelli, H. merlino, R. Martinez.,”Web Usage Mining using self Organized Maps.”,IJCSNS International Journal of Computer Science and Network Security Vol7 No 6 2007.
[24] Nidhi Raj, N.K.Singh,”Web Mining Techniques in The Area of the Web Personalization”, iraj, Volume-4, Issue-5, May 2017.
[25] Q. Han, X. Gao and W. Wu, Study on Web Mining Algorithm Based on Usage Mining, 2010.
[26] T. Bhatia, “Link analysis algorithms for web mining,” IJCST, vol. 2, Jun 2011.
[27] M. A. Preeti Chopra, “A survey on improving the efficiency of different web structure mining algorithms,” International Journal of Engineering and Advanced Technology (IJEAT), vol. 2, Feb 2013.