INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

WEB CONTENT MINING METHODS AND APPLICATIONS FOR INFORMATION EXTRACTION: A SURVEY

NARENDRA M. JATHE, DR. HEMANT S. MAHALLE

Department of Computer Science, Arts, Commerce and Science College, Kiran Nagar Amravati (M.H.)

Accepted Date: 15/03/2016; Published Date: 01/05/2016

Abstract*:* Web Mining is extracting information from the web resources and finding out of the ordinary patterns that can be useful from ever expanding database of WWW. One of the subfield of Web mining is Web Content Mining. Web data processing is the technique of handling high volume of data. Web mining provides high performance system to the users to search for the content/data and obtains information of a particular content/data by searching through the servers that contains the sources. Web content mining is simply an integration of data from various website sources and aggregate from various websites using web content mining techniques. This paper presents a survey on web content mining techniques used for mining and application of web content mining and also shows some of the promising techniques used for extraction of data from various websites using web content mining techniques.

Keywords: Web Content Mining, Information Extraction, web document types, Mining techniques and Attribute Extraction.

Corresponding Author: MR. NARENDRA M. JATHE

Access Online On:

www.ijpret.com

How to Cite This Article:

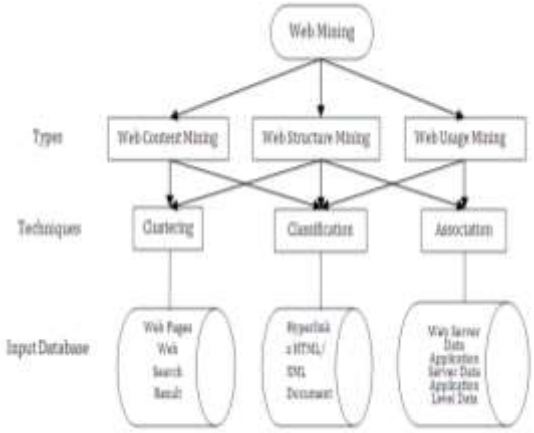
Narendra M. Jathe, IJPRET, 2016; Volume 4 (9): 794-801

PAPER-QR CODE

INTRODUCTION

Web is taking an important place in human's life and day by day it increases the number of information based on the expectations of the web users using it [1]. The World Wide Web has lot of information and continues to increase in size and difficulty. It is very extra special task to search relevant information from huge amount of data. The data used for web content mining includes both text and graphical data. Content mining is divided into two parts: webpage content mining and search result mining. In webpage content mining web is search via content. The search results content mining searches from the previous search result. When you search any specific key word or any web page, number of links or result is displayed. But all the data which is displayed on the web is not relevant. So efficiently and effectively retrieve required data on the Web is becoming a challenge [2]. Web mining is used to extract the web information that is needed by the users so that the necessary details can be fetched and utilized. Automation is everywhere and in every field to avoid the human work in creation of anything. Web mining utilizes the automatic way of information extraction from the World Wide Web according to the preferences [3].

The three categories used for mining the web are mentioned below in the following figure.



A. Web Content Mining

Web content mining is the process of mining, extracting and integration of useful data, information and knowledge from the web page content and web documents. It describes the discovery of useful information from the web documents. It contains the generation of wrappers. Wrapper is a set of extraction rules to extract the data from the web pages, this can done either manually of automatically. The collection of data to be integrated may contain images, texts, audios, videos and metadata and hyperlinks etc. Web content mining also distinguishes personal home pages with other web pages. This web content mining involves

document tree extraction, data classification, document categorization, data clustering and finally labeling the attributes for results. Research activities are going on in information retrieval methods, natural language processing and computer vision [2][3].

B. Web Structure Mining

The process of discovering structures information from the web documents are called as web structure mining. This mining can be performed either document level or hyperlink level. The hyperlinks provide clear navigation and point to the pages. This is used to retrieve the useful information in the form of structure. Hyperlink analysis can be done based on knowledge models, scope and properties of analysis and types of algorithms. The methods that are done in the web usage mining are Data cleaning, Transaction identification, Data integration, Transformation, Pattern Discovery, Pattern Analysis. This kind of mining emphasizes on the data which describes the structure of the content. It is classified into two types namely intra-page structure and inter-page structure. Intra-page structure means the existence of links within a page. No separate page will be opened in this case. Inter-page structure involves the connection of one page with the other page [3][2].

C. Web Usage Mining

Web usage mining is used to discover the interesting usage patterns form the usage data. This includes server data (IP address), Application server data (web logic), and Application level data (events). This is otherwise a Discovery of meaningful patterns from data generated by client-server transactions on one or more Web localities. The source database is access logs, referrer logs, agent logs, and client-side cookies [3]. It refers to the discovery of user access patterns from the web usage logs. It focuses on various data mining techniques to understand and analyze search patterns [2].

II. EMERGING TECHNIQUES FOR DATA EXTRACTION USING WEB CONTENT MINING

Based on the documents in the web the traditional methods are partitioned into four parts [4][5].

1. Unstructured Data Mining Technique
2. Structured Data Mining Technique
3. Semi-Structured Data Mining Techniques
4. Multimedia Data Mining Techniques

1. Unstructured Data Mining Technique: Content mining can be done on unstructured data such as text mining of unstructured data gives unknown information. Text mining is extraction of previously unknown information by extracting information from different text sources.

Content mining requires application of data mining and text mining techniques. Basic Content Mining is a type of text mining. Some of the techniques used in text mining are Information Extraction, Topic Tracking, Summarization, Categorization, Clustering and Information Visualization [6].

i. Information Extraction: Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi- structured machine-readable documents. Extracting information from unstructured data and converts into structured data. Pattern matching and transformation are used [1]. To extract information from unstructured data, pattern matching is used. It traces out the keyword and phrases and then finds out the connection of the keywords within the text. It utilizes feature extraction and key term indexing to build a graphical representation. This technique is very useful when there is large volume of text. Information extraction is the basis of many other techniques used for unstructured mining. Information extraction can be provided to KDD module because information extraction has to transform unstructured text to more structured data. First the information is mined from the extracted data and then using different types of rules, the missed out information are found out.IE that makes incorrect predictions on data are discarded [6].

ii. Topic Tracking: Tracks the topics searched by the user and predicts the documents and produce to the user that of interest. Prediction is used techniques. This technique checks the documents viewed by the user profiles. According to each user it predicts the other documents related to users interest. Disadvantage of topic tracking is that when we search for topics we may be provided with information which is not related to our interest [6].

iii. Summarization: Summarization is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document. Reduce the length of documents by maintaining the main points. Analyzing the semantics and interprets the meaning of words. It helps the user to decide whether they should read this topic or not. To understand the key points summarization tool search for headings and sub headings to find out the important points of that document. This tool also give the freedom to the user to select how much percentage of the total text they want extracted as summary. It can work along with other tools such as Topic tracking and categorization to summarize the document [7].

iv. Categorization: Categorization is the act of sorting and organizing things according to group, class, or, as you might expect, category. This technique is used to identify main themes by placing the documents into a predefined set of group. This technique counts the number of words in a document. It does not process the actual information. It decides the main topic from the counts. It ranks the document according to the topics [7].

Clustering: Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). Used to group the similar documents Grouping based on the properties are identified. Here in clustering, grouping is not done based on predefined topics. It is done based on fly. Same documents can appear in different group. As a result useful documents will not be omitted from the search results. Clustering helps the user to easily select the topic of interest. Clustering technology is useful in management information system [7].

vi. Information Visualization: To build a graphical representation to the user Feature extraction, indexing techniques are used. Through visualization documents having similarity are found out. Large textual materials are represented as visual hierarchy or maps where browsing facility is allowed. It helps the user to visually analyze the contents. User can interact with the graph by zooming, creating sub maps and scaling. This technique is useful to find out related topic from a very large amount of documents [8].

2. Structured Data Mining Technique: Structure data mining is most widely used in web content mining. Structured data is easier to extract as compare to unstructured data. Text mining of structured mining technique gives known information. The techniques which have been used for mining structured data are referred as Structured Data Mining Technique. Some of the techniques used in text mining are Web crawlers, Wrapper Generation, Page Content Mining, and Using OEM.

i. Web crawlers: A crawler is computer program that traverse/visits the hypertext structure in web sites and reads their pages and other information in order to create entries for a search engine index. The major search engines on the Web all have such a program, which is also known as a "spider" or a "bot". Traverse the hypertext structure of the web. There are two categories of Web Crawler such as: Internal and External Web Crawler. Internal Crawler crawls through internal pages of the Website which are returned by external crawler. External Crawler crawls goes through the unknown Website [8].

ii. Wrapper Generation: Set of information extraction rule to extract the useful data from web pages. Provides a lot of meta information lot of meta information Page ranking is used.

In Wrapper Generation, it provides information on the capability of sources. Web pages are already ranked by traditional search engines. According to the query web pages are retrieved by using the value of page rank. The sources are what query they will answer and the output types. The wrappers will also provide a variety of Meta information. E.g. Domains, statistics, index look up about the sources [8].

iii. Page Content Mining: Extracts the content of a page. Page ranking is used to display the results according to the rank. Page Content Mining is structured data mining technique which

works on the pages ranked by traditional search engines. By comparing page content rank it classifies the pages [9].

iv. Using OEM: Object Exchange Model. To understand the information structure of the web. Self-describing structure of the data is produced. Relevant information are extracted from semi-structured data and are embedded in a group of useful information and stored in Object Exchange Model (OEM). It helps the user to understand the information structure on the web more accurately. It is best suited for heterogeneous and dynamic environment [8].

3. Semi-Structured Data Mining Techniques: The use of Semi-structured data can be felt in the area involving raw data which does not have any fixed format. More and More data sets do not fit in the rigid relational model because of the Individual data items do not have the same structure completely. The techniques used for semi structured data mining are Object Exchange Model (OEM), Top down Extraction, and Web Data Extraction language [10].

i. Top Down Extraction: Complex objects of rich resources are converted into less complex objects. In top down extraction, it extracts complex objects from a set of rich web sources and converts into less complex objects until atomic objects have been extracted [8].

ii. Web Data Extraction Language: Converts web data to structured data and delivers to end users. Web data extraction language converts web data to structured data and delivers to end users. It stores data in the form of tables [8].

4. Multimedia Data Mining Techniques: It is a part of content mining where high level information and knowledge from large online multimedia sources. Multimedia data mining refers to the analysis of large amounts of multimedia information in order to find patterns or statistical relationships. Once data is collected, computer programs are used to analyze it and look for meaningful connections. This information is often used by governments to improve social systems. It can also be used in marketing to discover consumer habits. Some of the Multimedia

Data Mining Techniques are SKICAT, Color Histogram Matching, Multimedia Miner and Shot Boundary Detection [8].

i. SKICAT: Based on astronomical data analysis and cataloging system. SKICAT is a Successful Astronomical Data Analysis and Cataloging System that produces digital catalog of sky object. It uses machine learning technique to convert these objects to human usable classes. It uses machine learning technique to convert these objects to human usable classes. It integrates technique for image processing and data classification which helps to classify very large classification set.

ii. Color Histogram Matching: Find the correlation between the color components. Unwanted artifacts are removed using smoothing techniques. Color Histogram matching consists of Color

histogram equalization and Smoothing. Equalization tries to find out correlation between color components. The problem faced by equalization is sparse data problem which is the presence of unwanted artifacts in equalized images. This problem is solved by using smoothening.

iii. Multimedia Miner: Extraction of images. Videos for the feature extraction, and feature comparison for matching queries. Multimedia Miner Comprises of four major steps, Image excavator for extraction of image and Video's, a preprocessor for extraction of image features and they are stored in a database, A search kernel is used for matching queries with image and video available in the database. The discovery module performs image information mining routines to trace out the patterns in images.

Shot Boundary Detection: Automatic detection of boundaries. It is a technique in which automatically the boundaries are detected between shots in video [11].

III.APPLICATIONS OF WEB CONTENTMINING

Web content mining is used in various fields of large information maintenance. Cloud users need to extract the information from the cloud provided by web servers can utilize the web mining. Online shopping systems use the web mining to extract the information of a product and its specification through web mining. Web wide tracking is effectively done using web mining methodologies. Web communities can be maintained such as facebook. That is the users of same field of interest can be grouped and they can communicate through the network analyzed. Web page personalization now a days are very important to maintain the confidential information. Web mining is used for maintaining personalized data. Digital library performs automated citation indexing using web mining techniques. e-services include e-banking, search engines, on-line auctions, on-line knowledge management, social networking, e-learning, blog analysis, personalization and recommendation systems. This can be analyzed for the customers and enable provision to the customers based on their recommendations.

IV. CONCLUSIONS

Data mining techniques used for web information extraction are incredible system and recommended for the maintenance of highly confidential data. This is prosperous, most intelligent resource extractor, and useful to maintain the historical data. Vast amount of data is maintained by the web sources and can be clearly extracted by the web mining techniques when the techniques are used accurately based on the requirements of the users. Thus various Data mining techniques and web content mining tools are used to extract useful information or knowledge from web page contents.

This paper focuses on web content mining tools, techniques of web content mining and uses of Web Content Mining. By these techniques we can make our search of contents over the web faster and exact. The survey also discusses the techniques used for extracting information from

different types of data available in the internet and how this extracted data can be used for mining purposes. Users feel difficulty in finding desired information and deciding which information is relevant to them from general purpose search engines. Web content mining solves this problem and helps the users to fulfill their needs.

REFERENCES

1. Ananthi.J, "A Survey Web Content Mining Methods andApplications for Information Extraction from Online Shopping Sites", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 4091-4094.
2. Govind Murari Upadhyay, Kanika Dhingra, "Web Content Mining: Its Techniques and Uses", International Journal of Advanced Research in Computer Science and Software Engineering., Volume 3, Issue 11, November 2013, pp 610-613.
3. T. SunilKumar, Dr. K. Suvarchala, "A Study: Web Data Mining Challeneges and Application for Information Extraction", IOSR Journal of Computer Engineering (IOSRJCE), Vol 7, Issue3,Nov-Dec 2012,pp 24-29.
4. Faustina Johnson, Santosh Kumar Gupta, "Web Content Mining Techniques: A Survey", International Journal of Computer Applications (0975 – 888),Volume 47– No.11, June 2012,pp.44-50.
5. Govind Murari Upadhyay, Kanika Dhingra, "Web Content Mining: Its Techniques and Uses", International Journal of Advanced Research in Computer Science and Software Engineering, Volume3, Issue 11, November 2013,pp.610-613
6. Srivastava, J., Cooley, R., Deshpande, M., Tan, P. N. "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", 2000.
7. Fan, W., Wallace, L., Rich, S., Zhang, Z., "Tapping into the Power of Text Mining", Communications of the ACM – Privacy and Security in highly dynamic systems. Vol. 49, Issue 9, 2005.
8. ManjotKaur, Prof. NavjotKaur, "Web Content Mining Techniques: A Survey", International Journal of Computer Science and Technology, IJCST Vol. 4, Issue 2, April - June 2013pp 149-152
9. Nimgaonkar, S., Duppala, S.,"A Survey on Web Content Mining and extraction of Structured and Semi structured data", IJCA Journal, 2012.
10. Pol, K., Patil, N., Patankar, S., Das, C.,"A Survey on Web Content Mining and extraction of Structured and Semistructured Data", IEEE First International Conference on Emerging Trends in Engineering and Technology, pp. 543-546, 2008.
11. Smeaton, A. F., Over, P., Doherty, A. R.,"Video Shot Boundary Detection: Seven years of TRECVID Activity", Elsevier, Computer Vision and Image Understanding, Vol. 114, Issue 4, pp. 411-418, 2010.