Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 4, April 2017, pg.49 – 55

Web Content Mining Techniques and Tools

Andemariam Mebrahtu[1], Balu Srinivasulu[2]
[1]Lecturer, Computer Science, Eritrea Institute of Technology, Eritrea, North East Africa
[2]Lecturer, Computer Science, Eritrea Institute of Technology, Eritrea, North East Africa
[1] nwayandandat@gmail.com, [2] balusrini@gmail.com

Abstract — with the spectacularly and unpredictable growth of information available over the Internet, WWW has become a powerful platform for storage & retrieval of information. Due to the heterogeneity & unstructured nature of data on WWW, searching of information is becoming cumbersome & time consuming task. Web mining came as solution for the above problem. Web mining is the means of utilizing data mining methods to extract useful information from web. Web content mining is a sub division of web mining. In this paper, we first present the concepts of web mining, we then provide an overview of web mining techniques, and then we present an overview of different types of web content mining tools and conclude with the algorithms.

Keywords — Web mining, Web content mining, Web usage mining, Web content mining tools, and Web structure mining.

I. INTRODUCTION

The World Wide Web (WWW) is a popular and interactive medium with tremendous growth of amount of data or information available today. The World Wide Web is the collection of documents, text files, images, and other forms of data in structured, semi structured and unstructured form. It is also huge, diverse, and dynamic, hence raises the scalability. The primary aim of web mining is to extract useful information and knowledge from web. The web data store becomes the important source of information for many users in various domains. The web mining becomes the challenging task due to the heterogeneity and lack of structure in web resources. Because of these situations, the web users currently drowning in information and facing information overload [1]. Most of the web users could encounter the following problems, while interaction with the web;

A. Finding Appropriate Information

When a user wants to find specific information in the web, they input a simple keyword query. The query response will be the list of pages ranked depends on their similarity to the query. Though, today's search tools have some problems such as Low precision (due to the irrelevance of search results) and Low recall (inability to index all the information available).

B. Creation of New Knowledge from the Web

This problem is a data-triggered process whereas the previous is a query-triggered process. Here the web user has to extract potentially useful information from a collection of available contents.

C. Personalizing Data's

This is associated with the type and presentation of information, as it is likely that people differ in the contents and presentations they prefer while interacting.

D. Analyzing Individual User Preferences

This deals with the problem of encountering the needs of web users. This includes personalization of individual user, website design and management, customizing user information etc. The web becomes noisy if it contains various kinds of information. The web mining techniques can be used to solve those issues.

II. WEB MINING

Web Mining refers to the overall process of discovering potentially useful and previously unknown information or knowledge from the web data. Web Mining is used to capture relevant information, rating new knowledge out of the relevant data, learning about the different types of users. Web mining uses data mining techniques to automatically discover and extract information from web documents and services. Several other techniques like information retrieval, information extraction and machine learning have been used in the past to discover the new knowledge from the huge amount of data available in the web. These techniques have been compared with web mining, Information retrieval works by indexing text and then select useful information. Web mining comprises of two systems like information retrieval system and information extraction system. Web Mining is further classified in to three categories

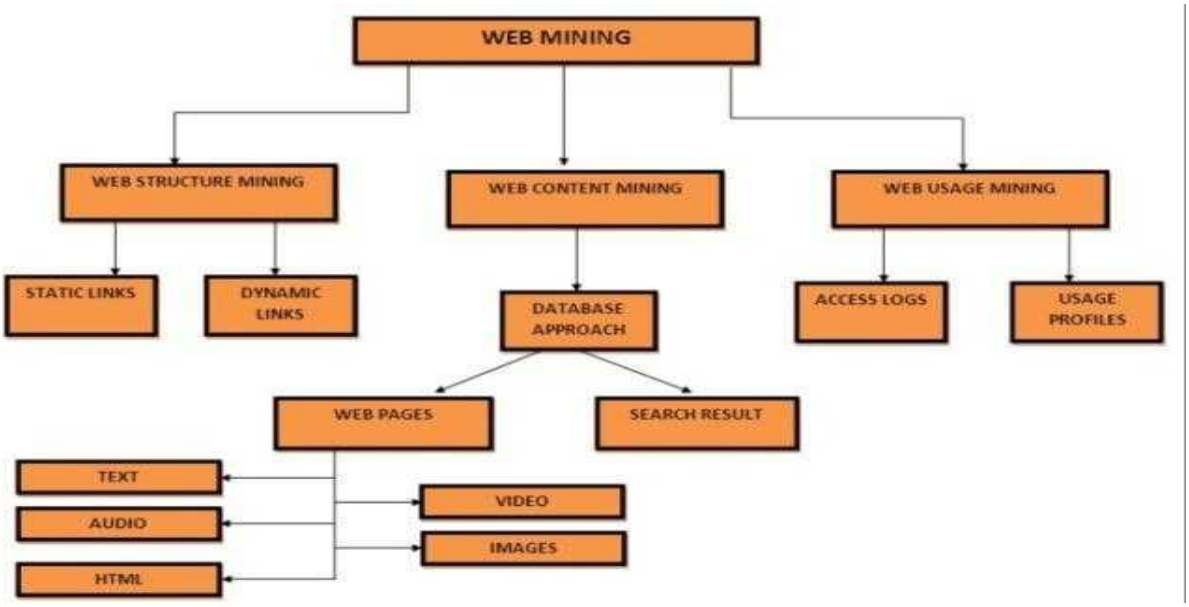

Fig.1 Web Mining Taxonomy

A. Web Content Mining

The seeable data on the web pages or any type of information which includes text, audio, video, images, HTML, XML is known as the content. To extract these types of data from different web pages comes under Web Content Mining. Web Content Mining comprises of excavating structured data, semi structured data or non structured data.

B. Web Usage Mining

Web Usage Mining is the litigation of eliciting any type of information from server logs [2]. It is the process of analyzing the curiosity of the users on the internet i.e. in what type of data they are interested for. For instance some users are interested in text type data or some other users are interested in audio, video or images. With the help of Web Usage Mining, we can study the behavior of the user. Using Web Usage Mining, users can get the different type of suggestions for which they are looking for .E.g. Property Search, Online Shopping sites for a particular product etc.

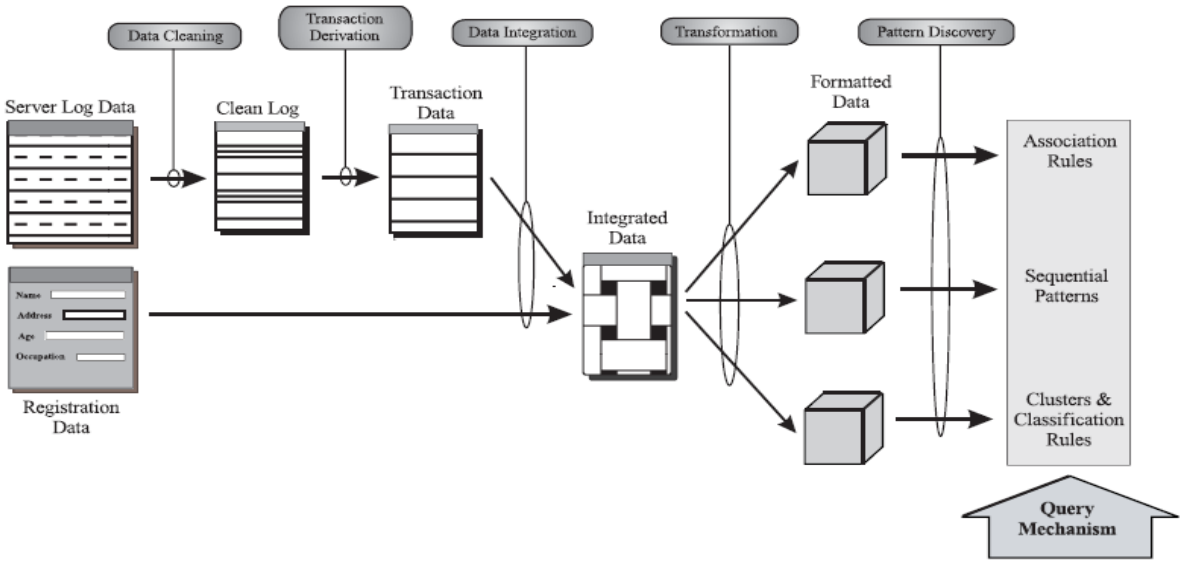

Fig.2 Web Usage Mining

C. Web Structure Mining

It is a tool practiced to discover the link between two or more webpage associated with information. The main intention of web structure mining is to take out the previously known relationships between the webpage.
It basically uses the graph theory with various nodes and the connection link to all the nodes. In the field of business or E-Commerce, a group of users i.e. clusters can be made for searching similar type of data on the web which results in improvement in several businesses very efficiently and increase in the production of sale.

III. WEB CONTENT MINING TECHNIQUES

Web Content mining has following approaches to mine data: unstructured mining, structured mining, semi-structured mining and multimedia mining.

A. Unstructured Data Mining

Text document is the form of unstructured data. Most of the data that is available on web is unstructured data. The research of applying data mining techniques to unstructured data is known as knowledge discovery in texts [3].

1) Information Extraction: To extract information from unstructured data that is present on web pattern matching is used. It traces the keywords and phrases and then finds out the connection of keywords within text. When large volume of text is there then the technique is very useful. Information extraction transforms unstructured text to more structured form. First, from extracted data the information is mined, then using different types of rules, the missed out information is found. Information extraction making incorrect predictions on data is discarded [4] [5].

2) Topic Tracking: This technique checks the documents viewed by the user and studies the user profile. It predicts the documents related to users interest. The topic tracking applied by yahoo, user give a keyword and if anything related to keyword pops then the user is informed about that. This technique can be applied by many fields. The two fields where it is used is medical field and education field. In medical field doctors easily come to know about the latest treatments. In education field it is used to find out the latest reference for research related work. The disadvantage of the technique is that when we search for our topic then it may provide us with information which is not related to our topic [4] [6].

3) Summarization: The technique is used to reduce the length of the document by maintaining the important points. It helps the user to decide whether to read the topic or not. The time taken by the technique to summarize the document is less than the time taken by

the user to read the first paragraph [4]. The summarization technique uses two methods that is the extractive method and the abstractive method. The extractive method selects a subset of phrases, sentences and words to form the summary from the original text. The abstractive method builds an internal semantic representation and then uses natural language generation technique to create the summary. This summary may contain words which are not present in the original document [6].

4) Categorization: This technique identifies the main theme by placing the documents in a predefined set of group. The technique counts the number of words in the document and this decides the main topic. According to the topic the rank is given to the document. The documents with majority contents on particular topic are given first rank. This technique helps in providing customer support to the industries and business [4] [5].

5) Clustering: The technique is used to group similar documents. In this grouping of documents is not done on the basis of predefined topics. It is done on fly basis. Some documents may appear in different group. As a result useful documents are not omitted from search results. This technique helps user to select the topic of interest [4].

6) Information Visualization: Visualization utilizes feature extraction and key term indexing. Documents having similarity are found out through visualization. Large textual materials are represented as visual maps or hierarchy where browsing facility is allowed. It helps in visually analyzing the content. The user can interact by scaling, zooming and creating sub maps of the graphs [3].

B. Structured Text Data Mining

Structured data are typically the data records retrieved from underlying database and displayed in the web pages. It can be displayed either as tables or forms. Data can be extracted from these sources using structured data extraction techniques. This can be helpful in making value aid services by collecting information from various sources e.g. customized Web information gathering, comparative shopping, meta-search. Following techniques are used for mining structured data:

1) Web Crawler: A web crawler is a relatively simple automated program, or script that methodically scans or "crawls" through Internet pages to create an index of the data it's looking for; these programs are usually made to be used only once, but they can be programmed for long-term usage as well. There are several uses for the program, perhaps the most popular being search engines using it to provide web surfers with relevant websites. Other users include linguists and market researchers, or anyone trying to search information from the Internet in an organized manner. Alternative names for a web crawler include web spider, web robot, bot, crawler, and automatic indexer. Crawler programs can be purchased on the Internet, or from many companies that sell computer software, and the programs can be downloaded to most computers. There are various uses for web crawlers, but essentially a web crawler may be used by anyone seeking to collect information out on the Internet. Search engines frequently use web crawlers to collect information about what is available on public web pages. Their primary purpose is to collect data so that when Internet surfers enter a search term on their site, they can quickly provide the surfer with relevant web sites. Linguists may use a web crawler to perform a textual analysis; that is, they may comb the Internet to determine what words are commonly used today. Market researchers may use a web crawler to determine and assess trends in a given market.

2) Page Content Mining: Page content mining is a technique that is used to extract structured data which works on the pages that are ranked by the traditional search engines. The pages are classified by comparing the page content rank [7].

3) Wrapper Generators: To facilitate effective search on the World Wide Web several Meta Search Engines have been formed which do not do the search themselves but take help of the available search engines to find the required information. Meta Search Engines are connected to search engines by the means of Wrappers. For every search engine connected to it, there is a wrapper which translates user's query in to native query language and format of the search engine. The wrapper also extracts the relevant information from the HTML result page of the search engine. [8]

IV. WEB CONTENT MINING TOOLS

Web content mining tools are software that helps to download the essential information for users as it collects appropriate and perfectly fitting information. Some of the tools are

A. Web Info Extractor (WIE)

This is a tool for data mining, extracting Web content, and web content Analysis and it can extract structured or unstructured data from Web page, reform into local file or save to database, place into Web server.

Features [9]:
- Facilitates to define extraction tools which enable no need of learning boring and complex template rules.
- Extraction of tabular and unstructured data to file or database
- Extraction of new content while updating and monitoring Web pages.
- Be able to deal with text, image and other link file
- Deal with Web page in all language
- Running multi-task at the similar time
- Facilitates recursive task definition

B. Mozenda

This is a tool to enable users to extract and manage Web data. The Users can setup agents that normally extract, store, and also publish data to multiple destinations. Previously information is in Mozenda systems, users can format, repurpose, and mash up the data to be used in other applications or as intelligence. There are two parts of Mozenda's scraper tool: Mozenda Web Console: Mozenda is a Web application that allows user to run agents, view all the results, organize those results, and export the data's extracted. Agent Builder**:** Agent Builder is a Windows application used to build data extraction project. Features [10]:
- Easy to use
- Platform independency (Runs only on Windows).
- Work place independence: Tuning the scraper, managing the scraping process and get scraped data from any computer connected to the Web.

C. Screen-Scrapper

This is a tool for extracting/mining information from web sites. It is used for searching a database, which interfaced with software to attain content mining needs. The programming languages such as Java, .NET, PHP, Visual Basic and Active Server Pages (ASP) can also be used to access screen scraper. Features [11]:
- Screen-scraper present a graphical interface allowing the user to allocate URL's, data elements to be extracted and scripting logic to traverse pages and work with mined data.
- Once these items have been created, from external languages such as .NET, Java, PHP, and ASP, the screen-scraper can be invoked.
- Facilitates scraping of information at cyclic intervals, the common purpose of this software and its services is to mine data on products and download them to a spreadsheet.
- A classifier example would be a metasearch engine where in a search query entered by a user is concurrently run on multiple web sites in real-time, after which the results are displayed in a single interface.

D. Web Content Extractor (WCE)

WCE is a powerful and easy to use data extraction tool for Web scraping, and data extraction from the Internet. This offers a friendly, wizard-driven interface that will help through the process of building a data extraction pattern and creating crawling rules in a simple point-and click manner. This tool permit users to extract data from various websites such as online stores & auctions, shopping, real estate, and economic sites, business directories, etc. The extracted data can be exported to a variety of formats, including Microsoft Excel (CSV), Access, TXT, HTML, XML, SQL & MySQL script and to any ODBC data source. Features [12]:
- Helps in the extraction or collection of market figures, product pricing data, or real estate data.
- Support users to extract the information about books including their titles, authors, descriptions, ISBNs, images, and prices, from online book sellers
- Helps users in automate extraction of auction information from auction sites.
- Help to Journalists extract news and articles from news sites.
- Helps people seeking job postings from online job websites finding a new job faster and with minimum inconveniences
- Extracting online information about vacation and holiday places, including their detailed descriptions from web sites.

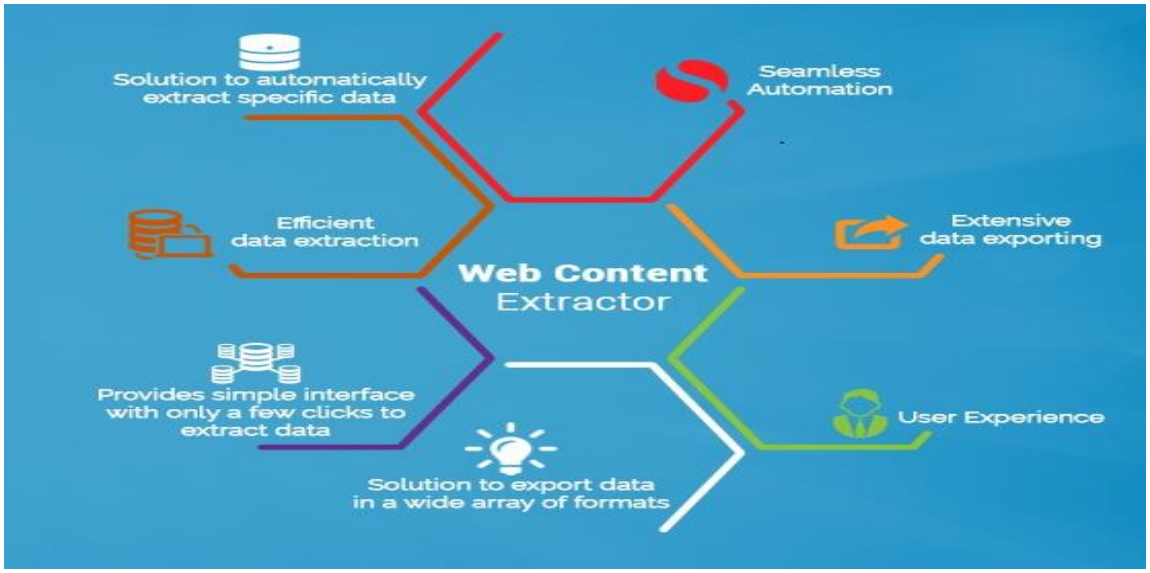

Fig.3 Web Content Extractor (WCE)

E. Automation Anywhere 6.1 (AA)

AA is a Web data extraction tool used in getting web data, screen scratch from Web pages or use it for Web mining. Features [13]:

- Automation Technology for rapid automation of complex tasks
- Recording keyboard and mouse or use point and click wizards to create automated tasks quickly.
- Web record and Web data extraction
- This has 305 plus actions were included Internet, conditional, loop, prompt, file management, database and system, automatic email notifications, task chaining, etc.

V. PERFORMANCE MEASURE

The performance metrics that is used to evaluate the performance of web content mining. The first main metrics is throughput. The throughput is the total time required to execute web content data [14]. The other performance metrics for web content mining evaluation are:

Cost Performance: It is measured as the ratio of throughput to cost of web content mining.

$$\text{Cost Performance} = \frac{\text{Throughput}}{\text{Cost}}$$

Scale up: It is capability of the system to manage more web content mining data with integrating more computers while maintaining the performance.

$$\text{Scale up} = \frac{\text{Throughput After}}{\text{Throughput Before}}$$

Latency: It is time to execute web content mining data set of operations.

$$\text{Latency} = \frac{1}{\text{Throughput}}$$

Durability: It is the ability of the system to maintain the information for extensive time period.

$$\text{Durability Ratio} = \frac{\text{Current Reads}}{\text{Total Reads}}$$

Concurrency: It is the ability of the system to provide a service to different users at the same time.

$$\text{Concurrency Ratio} = \frac{\text{Successful Operations}}{\text{Total Operations}}$$

There are many evaluation metrics and models to measure the performance of web content mining and data execution.

VI. CONCLUSION AND FUTURE RESEARCH

The web is the huge storage of network-accessible information, and knowledge. The web pages are continuously increasing in volume and complexity with time so it is going difficult to extract the valuable relevant information from internet. Thus several web mining techniques, methods and web content mining tools are applied to extract relevant useful information and knowledge from the web page contents. This paper reviews exploratory mining tools and techniques to mine the web contents in the internet. The analysis and theoretical review suggested the improvement of web mining algorithms. The parallelization process of huge volume of web data mining process can improve the performance in future. The parallelization process is the recommendation for future as the web data is continuously growing at rapid speed.

REFERENCES

[1] P. Maes. *Agents that reduce work and information overload.* Communications of the ACM, 37(7):30–40, 1994.

[2] M onik a Yad a v , Pra d eep Mi tt a l , " Web Mining: An Introduction ", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013 ISSN: 2277 128X.

[3] Johnson, Faustina, and Santosh Kumar Gupta. "Web Content Mining Techniques: A Survey."International Journal of Computer Applications (0975–888) Volume (2012)

[4] Sharma, Arvind Kumar, and P. C. Gupta. "Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining" International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 1 (2012)

[5] Srividya, M., D. Anandhi and M. I. Ahmed. "Web mining and its categories– a survey "International Journaof Engineering and Computer Science, IJECS 2.4 (2013)

[6] Deepti Sharda and Sonal Chawla "Web Content Mining Techniques: A Study."International Journal of Innovative Research in Technology & Science

[7] Johnson, Faustina, and Santosh Kumar Gupta. "Web Content Mining Techniques: A Survey."International Journal of Computer Applications (0975–888) Volume (2012)

[8] Boris Chidlovskii, Jon Ragetli, and Martin de Rijke "Automatic Wrapper Generation for Web Search Engines"

[9] Zhang, Q., Segall, R.S., *Web Mining: A Survey of Current Research, Techniques, and Software*, International Journal of Information Technology & Decision Making. Vol.7, No. 4, pp. 683-720. World Scientific Publishing Company (2008)

[10] Mozenda, *http://www.mozenda.com/web-mining-software*

[11] Web Content Extractor help. WCE, *http://www.newprosoft.com/web-content-extractor.htm*

[12] Screen-scraper, *http://www.screen-scraper.com*.

[13] Automation Anywhere Manual. AA, *http://www.automationanywhere.com*

[14] Amit Dutta, Sudipta Paria, Tanmoy Golui, Dipak K. Kole, "Structural analysis and regular expressions based noise elimination from web pages for web content mining" , IEE E International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1445– 1451, 2014.

AUTHORS BIBLIOGRAPHY

Mr. Andemariam Mebrahtu, Currently I am working as Lecturer and HOD in Department of Computer Science, Eritrea Institute of Technology, Asmara, Eritrea. I have sound experience in teaching, academic Administration activities and research in field of Computer Science. I have published a number of international journal papers related to the Computer Science. My area of interest includes Cloud Computing, Data Mining and Big Data Management.

Mr. Balu Srinivasulu currently I am working as a Lecturer in the Department of Computer Science, Eritrea Institute of Technology, Asmara, Eritrea. I have wide experience of teaching and research in field of Computer Science. I have published a number of international journal papers related to the Computer Science. My areas of research are Wireless Networks, Communication Networks, Big Data and Cloud Computing.

