International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:8, No:9, 2014

A Comparative Analysis of Different Web Content Mining Tools

T. Suresh Kumar, M. Arthanari, N. Shanthi

***Abstract*—**Nowadays, the Web has become one of the most pervasive platforms for information change and retrieval. It collects the suitable and perfectly fitting information from websites that one requires. Data mining is the form of extracting data's available in the internet. Web mining is one of the elements of data mining Technique, which relates to various research communities such as information recovery, folder managing system and simulated intellects. In this Paper we have discussed the concepts of Web mining. We contain generally focused on one of the categories of Web mining, specifically the Web Content Mining and its various farm duties. The mining tools are imperative to scanning the many images, text, and HTML documents and then, the result is used by the various search engines. We conclude by presenting a comparative table of these tools based on some pertinent criteria.

***Keywords*—**Data Mining, Web Mining, Web Content Mining, Mining Tools, Information retrieval.

I. Introduction

THE World Wide Web contains mass information, or is a mix up of lots of information like, texts, images, videos and others. Before knowing Web mining, we need to know Data Mining. Data mining is the process of extracting various information from a large collection of databases [1]. Like Oracle, SQL Server, Database etc. In the data mining process, a large number of data is created in structured forms like, tabular forms, files, and views. Data mining techniques process various information available where the data are available in tabular forms, files and views.

Web mining is the application of traditional data mining. With the growth of the text documents, text mining is becoming increasingly important and popular. Web mining used to capture the relevant information or data, and creating new knowledge from the relevant data.

Due to the rapid growth of the internet, sites appear, and disappear, contents need to be changed or modified in this competitive World. The World Wide Web is the one of the most desired option for any individual or an organization to search information [2]. The Web is huge and diverse, filled up with multimedia data and temporal issues respectively. Analysis and discovery of needed information from World Wide Web poses a phenomenal challenge to the researchers in this area. Such phenomena of collecting information by adopting data mining techniques are called Web Mining [3].

Many Organizations and companies use web mining technique to collect and share various information for their business development, and as well as for future. The primary aim for web mining is to extract the useful information and knowledge from the web. Now a day's web mining became the challenging task, owing to the heterogeneity and require of formation in web resources. Web data is typically enabled, distributed, semi-structured, time varying and high dimension. Almost 90% of the data are useless, often does not represent any relevant information that the user looking for. The web is dynamic. Information on the web changes continuously [4]. Keeping up among the changes and monitoring the changes are important issues. Web mining technique also raises an idea of data security of personal information available on the internet.

Web mining process can be explained as follows

- Finding Web Resources
- Select type of Information
- Generalize the Information
- Analysis the Information
- Extract Required Information

In the first step, the web resource and the source data should be located, and the source document is finalized. Selection of information is automated for processing information from web resources. The information will be generalized in the third process. The extracted information is validated and interpreted to find the patterns. The patterns are processed to represent information. After analyzing the information Web mining has many applications which help users extract useful information and make suitable decisions [7].

II. Web Mining

The main objective of web mining is to find out the useful information from page contents, Web hyperlinks and usage logs. The Web consists of a huge amount of data stored or collected around the Globe. Web mining helps to make the process of finding the needed information from a huge data like the web in an effective and effective way. It consists of various actions like, analytics, databases, processing, information retrieval, multimedia, etc.

Through by adopting various data mining techniques to retrieve valuable information from the web called data mining [3]. It refers to the use of collecting valuable information automatically from web documents through various data mining techniques [4]. The Web site is the one of the key

Mr.T.Suresh Kumar, Assistant Professor, is with the Department of Information Technology, K. S. Rangasamy College of Technology, Tamil Nadu, India (phone: 9865939876; e-mail:sureshitksr@gmail.com)

Mr.M.Arthanari, PG Scholar, is with the Department of Information Technology, K. S. Rangasamy College of Technology, Tamil Nadu, India (phone: 7708350866; e-mail: arthueaswer@gmail.com).

Dr.N.Shanthi, Professor and Dean, is with the Department of Computer Science Engineering, Nandha Engineering College, Tamil Nadu, India (e-mail: shanthimoorthi@yahoo.com).

communication channel for private individuals, trying to seek miscellaneous information, and not only for companies. Web page consists of various information like videos, images related to the web or other digital aspects [5].

The information's in the form of various research communities like artificial intelligence, database management systems and information collecting are well controlled from the ground principles [7]. For the business development activities, most of the companies use web mining technique to extort and share the valuable information [8]. How the web mining process is taking place is shown in the below Fig. 1. Simply web mining is to find out the previously unknown information or facts or finding out the preferred data from a huge database or from the web data. It is used to create new information from the existing data, customizing the information, understanding about the clients, personal users and several others [9]. Web mining is the application of various data mining techniques to find out the various patterns from the web. It refers to the process of updating the recent knowledge, or replacing the old information with the newer one.

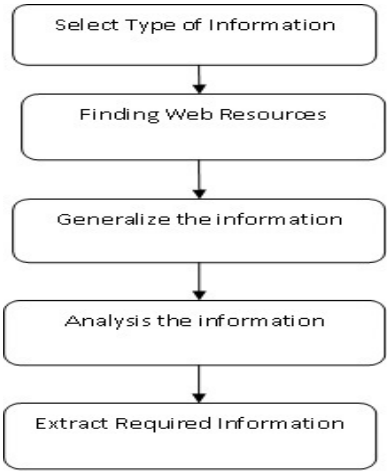

Fig. 1 Web mining Process

III. Web Content Mining

Web content mining is the process of collecting the required information from the text, image, audio or videos, already existed on the web. It is also named as web text mining. On the World Wide Web, text content is most widely researched topic. Web content mining will help to convert the broader form of the bulk data to a narrowed down useful information. Natural Language Processing and Information retrieval are the technologies that are commonly used in web content mining. In this section, we will see, what is the difference between Web content mining and Data Mining? [1].

It uses to extract the information from the web page; we can automatically categorize and gather web pages according their topics. These tasks are alike to those in usual data mining. It also helps to discover patterns in web pages to dig out valuable data like description of products, valuable forums, etc. The web content data are structured or pre-arranged data, such as data in the tables, shapeless information such as liberated texts, and semi-structured information such as HTML documents [2].

Web content mining although uses data mining techniques. But it differs, because web data are mostly shapeless or semi structured, while data mining deals with structured data. It relates to text mining, because most of the Web contents are texts. One of the main differences between these two, Web content mining and text mining is the Semi structural quality of the web [3].

Web structure mining is the mining of link structure concentrating on developing techniques to take benefits of the combined decision of web page quality which is available in the form of hyperlinks [4]. Web content mining or text mining is predominantly used in finding and tracking, clustering of web pages and sorting of web pages. In short, web content mining or text mining provides search engines to increase the flow of user clicks to websites, web pages of websites to solve their queries [5].

When it comes to technology, doing things manually will lead to a huge waste of time. Web mining helps to understand customer actions, helps to evaluate the performance of a website and the study done in web content mining will indirectly help to enhance business [6]. Scanning and pulling out of the graphs of a web page, pictures and text conclude the precise content to the query. Gopher, FTP and Usenet are the various services and data sources previously internet had. Now each and every service and data sources are easily reached via the web [7]. The consequential page, or documents are displayed by the search engine are according to the importance i.e. from highest significant information. It will help to reduce the wastage of time, and the pages displayed have links, so that users can easily go and get required data from there. There for unrelated information won't come unnecessarily. It removes the irritation and increase the navigation of data on the web.

Information Retrieval (IR), and Database (DB) views, are the two different points we could differentiate the research in web content mining. IR is mainly to improve the information finding or filtering the data to the user. Whereas, web content mining in sight of the database is that it tries to model data on the web and put together them in more sophisticated queries [7]. A Web page may consist of text, audio, video such as lists or tables, a Web page was designed to communicate to the users. Issues addressed in text mining are, extract organization patterns, area invention, classification of Web Pages and clustering of web documents [8].

IV. Web Content Mining Approach

Mainly there are two types of approaches for web content mining, namely,

1. Unstructured text mining approach
2. Semi-Structured and Structured mining approach.

A. Unstructured Text Data Mining (Text Mining)

Web content information is maximum in the form of unstructured text data. The study about applying data mining

techniques to unstructured or shapeless text is termed knowledge discovery in texts (KDT). Also named as or text mining, or text data mining [3], [6].

Some techniques are used in text mining are

- Information Extraction
- Topic Tracking
- Summarization
- Categorization
- Clustering
- Information Visualization.

B. Semi-Structured and Structured Data Mining

Structured data is also easier to extract compared to unstructured texts. For the web and database communities, semi-structured data is a meeting point, earlier it deals with document and later with data. Because of representing their host pages on the web, structured data are frequently very important. Variations on the Object Exchange Model (OEM), is one of the growing representations for semi-structured data (such as XML).

In OEM, information is in the form of atomic or compound objects: atomic objects may be integers or strings; compound objects refer to other objects through labeled edges [3].

The techniques used for mining structured data are

- Web Crawler,
- Wrapper Generation,
- Page content Mining.

V. Different Types of Tools

A. Web Info Extractor

This tool helps in pulling out web data, extracting web content and monitoring content update. It helps the people those who seeking for a job, resolve able to extract job posting from online job portal. Difficult templates are not required to be defined. It can extract structured and unstructured data from web page and transforms into local file or save the database, place into the server [1], [3].

Features:

- Facilitates to describe extraction tools.
- Tabular and unstructured data file or database extraction.
- Web pages are updated and monitoring the new content file.
- It supports Web pages in all languages.
- It is a successively multi - task at the same time.
- Unicode maintains can process web page in all languages

B. Mozenda

This tool allows users to extract the data, and manage efficiently. Users can set up agents, view & organize results, and exports publish data extracted [1]. Mozenda now supports logins, paging throughout, list of results, frames, AJAX with other difficult web sites. Once information is in the Mozenda, system users can repurpose, format, and mash up the information to be used in other online/offline applications or as intelligence [3]. Mainly there are two types of Mozenda Scraper tool, namely

- ✓ Mozenda Web Console
- ✓ Agent Builder

Mozenda Web console is an application is permitted the users to run agents; view & organize results and export publish data extracted.

Users can make the agent permanent those who are regularly extracted store and circulate information to various destinations.

Features:

- Working Environment independence.
- Platform independence.
- User Convenience to extract the data

Agent Builder is a Windows application used to build data extraction project.

Features:

- Easy to use
- Runs only on Windows (Platform Independent)
- Working Place is independence.

C. Screen-Scraper

It is another tool for extracting/mining information from web sites. It allows taking out the content from the web like, searching a database, SQL server or SQL database. The programming languages such as Java, .NET, PHP, Visual Basic and Active Server Pages (ASP) can also be used to access screen scraper [1], [3]. Screen-scraping is a tool for filtering information from web sites which can be used in other situations. Mine data on products and downloading them to a spreadsheet is the one of the most regular usage of this software [10].

Features:

- Created by external languages like, .NET, Java, PHP, and Active Server Pages
- Screen scraper can be accessed by some programming language.
- Download mine data products to a separate spreadsheet.

D. Web Content Extractor (WCE)

For the web scraping, it is considered as the most powerful and easy to use data mining tools [1]. This tool permits users to take out information from various websites such as online supplies, online public sale, shopping sites, valid domain sites, economic site, trade directory, etc. [2]. The collected information can be exported to a variety of formats, including Microsoft Excel (CSV), Access, TXT, HTML, XML, SQL script, My SQL script and to any ODBC data source [9], [11].

Features:

- It helps to extort/gather the primate data, product pricing data, or valid domain records.
- It helps users to mining the information concerning books, including their titles, authors, metaphors, ISBNs, descriptions, and prices, from online booksellers.
- Assists users in automate mining of auction information beginning auction sites.
- Assists to the media mining news and articles from report sites.

- It helps people looking for a job extract career postings from online job websites. Find a latest job faster and with minimum inconveniences.
- This tool helps businessmen mine and assemble the primate data, invention pricing data, or actual domain data.
- It helps users in computerize mining of auction information beginning auction sites.

E. Automation Anywhere

Automation anywhere is a web data extraction tool used for retrieving web data easily, screen scrape beginning web pages or utilize it for web mining. Intelligent automation is used for business and IT tasks [1], [2].

The Intelligent Automation Software automates and schedule business process and IT tasks an easier way [3].

Features:

- Automation Technology is used for rapid computerization of complex tasks.
- Demo the ivories and mouse or use indicate and click wizards to make automatic errands rapid.
- Web record and Web data extraction.
- This has 305 plus actions were built-in: Internet, provisional, disk, timely, folder administration, file and arrangement etc.
- It creating automation errands takes only some minutes, record keyboard and mouse strokes, or use easy point-and-click wizards.

F. HIT

HIT stands for Hyperlink Image Text mining. It is a Java based desktop application used to access the web URL in easy way. It gives the source code for the given URL. It is used to my web information only from the client side, and not from the server side. It also provides users easily access the hyperlinks, images on the web page in addition it helps to find the count of a particular word in a web page [9].

The HIT tool window shows the source code of the web page when the user clicks “Load Source code Button” and the user can count the number of words by clicking “Count Text Button” [9], [12].

Features:

- User can easily access any web source code by single clicking the URL
- User friendly to count a particular word in the given source code.

VI. Summarization of Web Content Mining Tools

The following table demonstrates the comparison of different tools in web mining based on their records data, extract structure and unstructured data. The main aim of this comparative study is to illustrate the usage paradigm and the awareness of the tools in web mining.

TABLE I
Comparison of Web Content Mining Tools

Name of the Tool	Records the Data	Extract Structured Data	Extract Unstructured Data	User Friendly
Automation Anywhere	Yes	Yes	Yes	Yes
Web Info Extractor	No	Yes	Yes	Yes
Web Content Extractor	No	Yes	Yes	Not For Unstructured Data
Screen Scraper	No	Yes	Yes	No
Mozenda	No	Yes	Yes	Yes

VII. Conclusion

In this paper, we presented a list of the accessible Web Content Mining Tools. The significance of web mining continues to increase due to the increasing propensity of web documents. Web mining ranks the various websites which help the organizations to find the user’s behavior, needs, preferences, etc. So that organizations can promote their products properly and to gain maximum profit The mining tools are imperative to scanning the many imagery data, HTML documents and text provided on Web pages. . We are currently working to design and implement a Web mining system based on multi-agents technology. We imagine that such system reduces the information overload and Search depth.

References

[1] *C* Abdelhakim Herrouz, Chabane Khentout, Mahieddine Djoudi, “Overview of Web Content Mining Tools” The International Journal of Engineering And Science (IJES) Volume 2 Issue 6 Pages 2013.
[2] Raymond Kosala, Hendrik Blockeel, “Web Mining Research: A Survey” ACM SIGKDD, July 2010. Volume 2, issue -1, page-1.
[3] V. Bharanipriya, V. Kamakshi Prasad, “Web content mining tools: A Comparative study” International Journal of Information Technology and Knowledge Management, January-June 2011, Volume 4, No. 1, pp. 211-215.
[4] Dragos Arotaritei, Sushmita Mitra,“Web mining: a survey in the fuzzy framework” Fuzzy Sets and Systems148 (2004)519.
[5] Arvind Kumar Sharma, 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, Issue 8, October 2012.
[6] R. Malarvizhi, K. Saraswathi, “Web Content Mining Techniques, Tools & Algorithms – A Comprehensive Study, ” International Journal of Computer Trends and Technology (IJCTT) – volume 4 Issue 8–August 2013.
[7] Monika Pathak, Sukhdev Singh, “A Comparative Study of Various Issues in Web Mining” International Journal of IT, Engineering and Applied Sciences Research (IJIEASR) Volume 3, No. 2, February 2014.
[8] T. S. Anushya Davy, M. Selvanayaki, “An Overview Of Web Content, Mining and its Techniques” International Journal of Advanced Science, Engineering and Technology Vol.3 Issue.1 No.10, March-2014, 48-53.
[9] Tripurari Pujan Pratap Singh, Dr. Anurag Seetha, K. K. Pandey, “HIT: Web Content Mining Tool” International Journal of Electronic Communication and Computer Engineering Volume 3, Issue 6.
[10] T. Shanmugapriya, P. Kiruthika, “Survey on Web Content, Mining and Its Tools” International Journal of Science, Engineering and Research (IJSER) Volume 2 Issue 8, August 2014.
[11] Nirali N. Madhak, Shahida G. Chauhan, Chintan R. Varnagar, “Understanding the Scope of Web Mining -Comprehensive Study”

National Conference on Emerging Trends in Computer & Electrical Engineering.
[12] M. Karpagam, R. Sasikala, "Analysis of Web Content Mining Tools" International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 12, December 2013.