Al-Qadisiyah Journal of Pure Science

Volume 25 | Number 2 Article 13

4-7-2020

# A STUDYING OF WEBCONTENT MINING TOOLS

Rasha Hani Salman
*Informatics Institute of Higher Studies, Baghdad, Iraq*, Rashahany609@gmail.com

Mahmood Zaki
*Mustansiriyah University,College of Engineering,Baghdad –Iraq*, drmzaali@uomustansiriyah.edu.iq

Nadia A. Shiltag
*University of Baghdad,College of Engineering,Baghdad –Iraq*, nadia.alijamali@coeng.uobaghdad.edu.iq

## Recommended Citation

Salman, Rasha Hani; Zaki, Mahmood; and Shiltag, Nadia A. (2020) "A STUDYING OF WEBCONTENT MINING TOOLS," *Al-Qadisiyah Journal of Pure Science*: Vol. 25: No. 2, Article 13.
DOI: 10.29350/2411-3514.1202
Available at: https://qjps.researchcommons.org/home/vol25/iss2/13

QJPS
Al-Qadisiyah Journal of Pure Science

Al-Qadisiyah Journal of Pure Science

ISSN(Printed): 1997-2490 ISSN(Online): 2411-3514

DOI: /10.29350/jops.

http://qu.edu.iq/journalsc/index.php/JOPS

# A STUDYING OF WEB CONTENT MINING TOOLS

***Authors Names***

Rasha Hani Salman[a]
Mahmood Zaki[b]
Nadia A. Shiltag [c]

***Article History***

Received on: 19/02/2020
Revised on: 09/03/2020
Accepted on: 07/04/2020

***Keywords:***

Data mining, web mining, web content mining, KNIME, Rapid-Miner, MOZENDA

**DOI** //doi.org/10.29350/ jops.2020.25.2.1067

**ABSTRACT**

The World Wide Web has become crowded with different data, which makes data mining a cumbersome and tiring process. Therefore, the web uses various information mining strategies to extract useful information from the web. Among these strategies is web content mining tools that are used to collect, sort, classify and provide the best data that can be accessed by the user. Web content mining tools are necessary to scan the HTML documents, images, and texts, the results are provided for the search engines. It can assist search engines in providing productive results of each search in order of their relevance. This paper presents an introduction to the concepts related to data mining and web mining, web content mining techniques, and the study of different web content mining tools by creating a comparative table of these tools based on some pertinent criteria.

## 1. Introduction

Data mining portrayed by Encyclopedia of Britannica it classified "knowledge discovery " in databases. In the field of computer science, it is the way for finding patterns and associations with regard to a major measure of big data. This field gathers tools from data statistics and artificial intelligence. With the management of the database to analyze big digital groups known as data sets. Data mining is broadly utilized in business (protection, banking, retail), science explores (space science, prescription medicine), and government security (discovering terrorists and criminals) [20][17]. Web mining is the application of data mining techniques which is unstructured or semi-structured data and it automatically discovers and extracts potentially useful and previously unknown information or knowledge from the web[9].

[a] Informatics Institute of Higher Studies, Baghdad, Iraq, E-Mail: Rashahany609@gmail.com
[b] Mustansiriyah University, College of Engineering ,Baghdad – Iraq, E-Mail: drmzaali@uomustansiriyah.edu.iq
[c] University of Baghdad, College of Engineering ,Baghdad – Iraq, E-Mail: nadia.alijamali@coeng.uobaghdad.edu.iq

The significant web mining applications are website design, web search, search engines, information retrieval, network management, e-commerce, business, and artificial intelligence, web market places, and web communities. The web mining process includes four important steps, these are resource finding, data selection, pre-processing generalization, and analysis resource finding. It is the process used to extract the data either from online or offline text resources. In data selection and preprocessing steps, specific information from retrieved web sources is automatically selected and pre-processed. During generalization, data mining and machine learning techniques are used to discover general patterns from individual web sites as well as across multiple sites. Validation and interpretation of the mined patterns are done in the analysis step. Web mining is classified into three different categories, i.e. web content mining, web structure mining, and web usage mining[24]. This paper consists of six sections.Section1 describes the introduction of data mining and web mining, while section2 exposes web mining categories. Section3 and section4 describe web content mining techniques and tools. Section5 describes a comparison table of Web Content Mining tools. Finally, chapter 6 conclusion and future work.

## 2. Web Mining Categories:

Web mining is comprehensively arranged into three various categories, as indicated by the sort of information to be mined [24]:

- Web Structure Mining (WSM).
- Web Usage Mining (WUM).
- Web Content Mining(WCM).

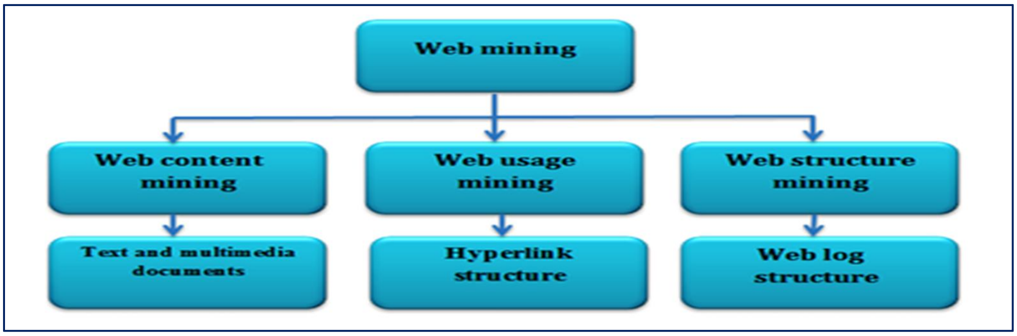

Fig.1: Web Mining Taxonomy

### 2.1 Web Structure Mining :

A web structure mining is a study for data related to the structure of a particular website. It consists of a web graph that contains web pages or web reports as nodes and hyperlinks as edges that connect two connected pages. Web structure mining is to extract some interesting web graph patterns like co-citation, social choice, complete graphs, etc. The site page is ranked on different points and the web page is chosen to be included in the page group. Web structure mining can be done either at the page level or between the page level. The hyperlink that overlaps with a different part of the same page is called the hyperlink inside the page. It is a document structure level [9].

## 2.2 Web Usage Mining :

Web usage mining is also defined as web log mining which is utilized to analyze the behavior of users on the web [4]. There are two kinds of tracking, first is general tracking and the other is a customizable usage tracking [24]. The general access tracking is utilized to foresee the conduct of the user on the web and it recognizes the user when the user connects with the web. It can save the information naturally when the web server log and applications log [18].

## 2.3 Web Content Mining :

Web content mining can be defined as a significant process of extracting the beneficial and necessary information from web pages. It is related to text mining since the vast majority of the web content is text-based. Web content mining is the semi-structured nature of the web. It has two types: Those that directly mine the content of documents and those that improve on the content search of other tools like search engines. Content mining is used to examine data collected by search engines and web spiders. The technologies that are normally used in Web Content Mining are Natural Language Processing (NLP) and Information Retrieval (IR) [7]. Two methodologies used in the web content mining are database and agent-based approach. The database approach helps in retrieving the semi-structured data from web documents [16][22]. The three kinds of agents are customized web agents, information filtering Categorizing agent and intelligent search agents [16]. Customized web agents try to find a web page on the user profile. Information filtering categorizing agents reduce the user's time and effort in locating the relevant document through the specialized domain knowledge it possesses. The filtering agent filters out irrelevant incoming documents and presents to the user only those documents which match the user's interest. Automatically, Intelligent search agents discover information according to a particular query utilizing user-profiles [16][22] .

## 3. Web Content Mining Techniques

Web content Mining has the accompanying ways to deal with extract information[9]:

(1) Unstructured text mining.

(2) Structured data mining.

(3)Semi-organized data mining.

(4)Multimedia data mining as shown in Figure 2

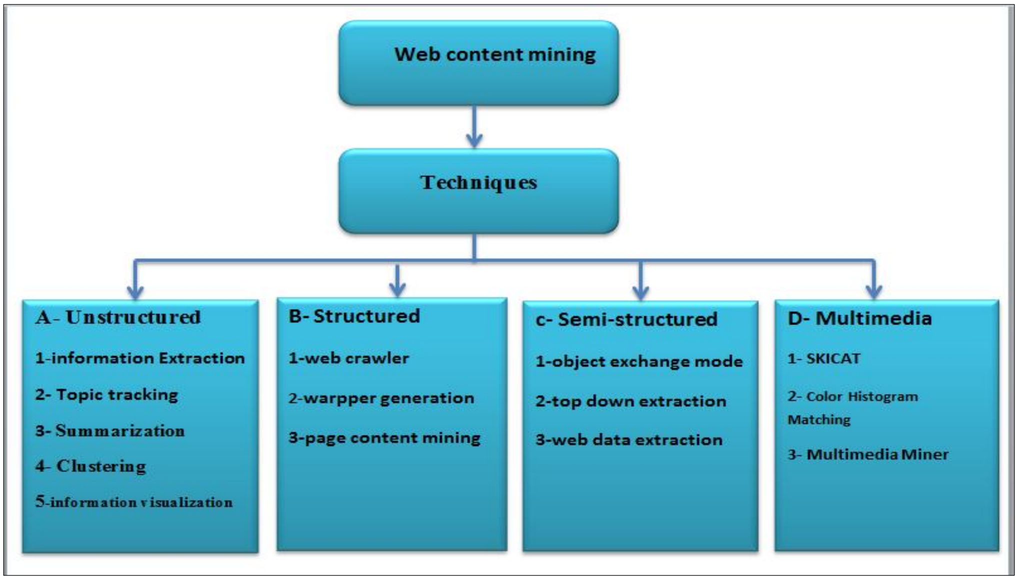

**Fig2: Web Content Mining techniques.**

### 3.1 Unstructured Text Mining:

Web content data is much of unstructured text data. The research around applying data mining techniques to unstructured text is termed knowledge discovery in texts (KDT), or text data mining, or text mining. Hence, one could consider text mining as an instance of web content mining. To provide effectively exploitable results, preprocessing steps for any structured data is done using information extraction, text categorization, or applying natural language processing ( NLP ) techniques[6].

### A. Information Extraction:

The information can be extracted from unstructured data, by utilizing pattern coordinating. It tracks the keyword and expressions and afterward finds the connection of the keyword inside the content[30]. This technique is very useful when there is a large text size. Information mining is the basis of many other technologies used in unorganized mining and can be provided to discover the knowledge of the database unit (KDD) because information mining should convert unorganized text into more structured data. First, the information is extracted from the extracted data, then by using different types of rules, the lost information is found. Information that makes incorrect predictions is extracted from the data[29].

### B. Topic Tracking :

This technology examines the documents the user viewed and studies user profiles, according to each user who predicts other documents related to the user's interest. The topic that tracks the Yahoo application, the user can give a keyword and if anything related to the keyword the user will be notified. The same can be applied to unstructured mining data. Example of tracking a topic if we choose the competitor's name, the name will appear in the news at any time and this information will be the company that passed. The feature tracking application can be in two areas: the medical and education field [29].

### C. Summarization:

Summarization is used to reduce the length of the document by maintaining the main points. It helps the user to decide whether they should read this 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. The challenge in summarization is to teach software to analyze semantics and to interpret the meaning. This software statistically weighs the sentence and then extracts important sentences from the document. To understand the key points summarization tool search for headings and subheadings to find out the important points of that document. This tool also gives the freedom to the user to select how much percentage of the total text they want to be extracted as a summary. It can work along with other tools such as Topic tracking and categorization to summarize the document. An example of text Summarization is Microsoft word's AutoSummarize [10].

### D. Categorization:

This technology places the documents in a pre-defined set of the set. Then, it counts the number of words in the document and this decides the main topic. According to the topic, the rank is awarded to the document. The documents that contain the majority contents on a specific topic are given first order. This technology helps provide customer support to industries and businesses[ 2][14].

### E. Clustering:

This technique has been used to group similar documents. Here in clustering, the grouping is not done based on predefined topics. It is done based on fly. The same documents can appear in a different group. As a result, useful documents will not be omitted from the search results. The clustering technique helps the user to easily select the topic of interest. Clustering technology has been used in Management Information Systems(MIS)[19].

### F. visualization's Information:

This technique is used to enhance user intelligence, visual presentations of abstract data are studied in this approach. A large amount of textual content is analyzed visually by the user in maps visualizing information. This process is vigilant when the user requests to see very large documents. By creating sub-maps, the users can interact with zoom in and out and scale[10] .

## 3.2 Structured Data Mining Techniques:

The systems are used to extract organized information from pages. Information in the Shape of tables, lists, and organized information tree. Organized information is simple to extract when contrasted with unstructured information[11].

### A. Web Crawlers:

Crawlers are soft wares that navigate the hypertext-structure on the web. It can be used by the user to collect data from the web. Web engines use crawlers regularly to assemble data about what is on open site pages. For information extraction web crawlers make use of methods like breadth-first search, particle swarm optimization, genetic algorithm, and other soft computing methods. And they also navigate through hyperlinks of web structure to extract knowledge [19].

### B .Wrapper Generation:

Wrapper generator is supplied data the information is provided by the wrapper generator on the capability of sources. Web pages are ranked by traditional search engines. by using the page rank value the web pages are retrieved according to the query[10].

### C. Page Content Mining:

This is a strategy used to extract organized information at the pages that are ordered by conventional web crawlers. The pages are ordered by looking at the page content position [10].

## 3.3 Semi-Structured Data Mining:

Semi-structured data are not full and grammatical text. Semi-structured data is hierarchical structured. Semi-structured data do not have a predefined structure. There are several techniques to extract semi-structured data; some of them are natural language processing (NLP) techniques, wrapper generation, ontology. To extract such data common approach is to build a specific grammar which details the surrounding of each piece of data to extract[19].

### A. Object Exchange Model (OEM):

The relevant information is extracted from semi-structured and is collected in a group of useful information and then stored in Object Exchange Model (OEM). This helps the user to accurately understand the structure of the information that is available on the web. The main feature of the model is that it is self-describing, i.e. there is no need to describe the structure of an object in advance[23].

### B. Top-down Extraction

This procedure helps with extract complex items from well off-sources of the web and breaking down them to things lesser complicated Until the atomic elements are separated [2].

### C. Web Data Extraction Language:

This technique helps in converting web data to structured data and then delivers this data to end-users. The data is stored in the form of tables[10].

## 3.4 Multimedia Data Mining:

Multimedia data mining is the process of finding interesting patterns from media data such as video, audio, text, and images that are not accessible using queries[23].

### A. SKICAT

SKICAT is a successful astronomical data analysis and cataloging system that produces a digital catalog of sky objects. It uses machine learning techniques to convert objects to human usable classes. It integrates technique for image processing and data classification which helps to classify very large classification sets[15] .

### B. Color Histogram Matching

This technique is made up of shading histogram equation and smoothing. Equation attempts to discover the relationship between parts of the color. Equalization is a confronted issue that is the presence of unneeded ancient rarities in adjusted pictures. Smoothening is utilizing to take care of this issue [23].

### C. Multimedia Miner

Multimedia miner consists of four major steps. 1- Image excavator for extraction of images and videos, 2- a preprocessor for extraction of image features and are stored in a database. 3- A search kernel is used for matching queries with images and videos that are available in the database. 4- The discovery modules mine image information to trace out the patterns in the image[2] .

## 4. A Study of Web Content Mining Tools

Several types of research studies have covered several web content tools. In the current studies, a number of these tools are presented taking into consideration extract important features for each tool.

### 4.1 Web Info Extractor (WIE)

This tool is used for mining web data and content retrieval and it is a very useful tool. It can retrieve unstructured or structured data from web page reorganize into local file or save to database, place it into a web server. Difficult template rules are not required to be defined, can the user browse to the web page and click what user desire to describe the retrieval task, and run it, or let it run automatically[28] .

### Feature

• Web Info Extractor is easy to define extraction task and no need to learn boring and dense template rules, monitor web pages and retrieve new content when update, this tool is Unicode support and can process web page in all languages, support recursive task (child task) definition. [28].

• This tool can deal with text, image, and another linked file. and it can deal with the web page in all languages. It can be running multi-task at the same time[1].

• It helps individuals looking for work to find a suitable job online[27] .

• It is a commercial tool that uses different formats such as excel (CVS) and text (TXT), the loading of this tool from the web site can be time-consuming, this tool cannot possible to record the data and it can support windows 2000/ XP/Vista OS[21].

### 4.2 SCREEN-SCRAPER:

A graphical interface is provided by the screen-scraper allowing you to designate URLs, data elements to be extracted and scripting logic to traverse pages and work with extracted data. Once these items have been created from external languages such as NET, Java, PHP, and active server pages, screen-scraper can be invoked. This also facilitates the scraping of information at periodic intervals[28].

### Feature

• SCREEN-SCRAPER searches a database, SQL Server or SQL database, which interfaces with the software to achieve content mining requirements. This is a tool for filtering information from web sites[10][27].

• The latest SCREEN-SCRAPER supply the information in HTML, thus it can be able to access it with a browse [3].

• It is a commercial tool, easy to automate the important website (fill in the open link form, etc.), and constantly monitor the web page to detect any change. This tool cannot record data, and it supported Windows 2000 / XP / Vista OS[21].

• SCREEN-SCRAPER using language such as Java, Python, NET, VB, and supported Linux [13].

### 4.3 MOZENDA

This tool enables users to extract and manage web data. Users can set up agents that routinely extract, store, and publish data to multiple destinations. the users of this tool can format, recreate, and mashup with the data as per the requirement so it can be used in other applications or as needed [9] .

### Feature

• MOZENDA web console: It is a web application that allows users to run agents, view and organize results, and export publishes data extracted and can be used easily, its Platform is independence[1].

• MOZENDA now supports logins, paging throughout, list of results, frames, AJAX with other difficult web sites[27].

### 4.4 Web Content Extractor (WCE):

This tool permits users to take out information from various websites such as online supplies, online public sales, shopping sites, economic sites, trade directories, etc. The collected information can be exported to a variety of formats [2].

### Feature

• It provides the user a friendly, wizard-driven interface [2].

• It helps users to mine the information concerning books, including their titles, authors, metaphors, descriptions, and prices, from online booksellers, helps to gather the primate data, product pricing data, or domain records assists users in automating mining of auction information sites. and helps in media mining news and articles from report sites [26].

• It is a powerful and easy tool for web scraping, data mining and data retrieval [31].

• Web content extractor is a commercial tool, used for extraction of URL, phone, Meta tag, e-mail, address, etc. Highly automated extensive training required [21].

### 4.5 Automation Anywhere:

The Intelligent automation software, 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 information technology (IT) tasks[1].

#### Feature

• This tool is smart's unique automation technology that quickly automates complex tasks, automation tasks take few minutes to record a keyboard and mouse stroke, and can be used easily, this tool can be Distributes tasks to multiple computers, it is used for task scheduling and automatic Tasks are scheduled at any time even when the computer is turned on or off locked [28].

• Automation Anywhere is commercial tools, very fast, highly automated, extensive training required[21].

### 4.6 SCRAPY

SCRAPY is a very useful tool to learn for any data professional. SCRAPY is free and open-source software written in Python for extracting data from websites. It is an application framework for extracting structured and crawling data used for like data mining applications and information processing [5].

#### Feature

This tool is useful for testing a web page, helps in monitoring, This tool is considered a difficult tool to use compared to other tools and it is a non-commercial tool[21].

### 4.7 WEKA

WEKA is a group of machine learning algorithms including pre-processing on data, classification, clustering, and association rule extraction for data mining applications. These algorithms can either be applied directly to a data set or can be called from your own Java code. The WEKA (pronounced Weh-Kuh) workbench contains a collection of many tools for visualization and algorithms for analytics of data and predictive modeling, together with graphical user interfaces for easy access to this functionality [12].

#### Feature:

• This tool is suitable for developing new machine learning schemes, WEKA loads data file in formats of (ARFF, CSV, C4.5binary). This tool is open-source, Free, extensible, and can be integrated into other java packages [21].

• WEKA has three graphical user interfaces such as the explorer for exploratory data analysis to support preprocessing, attribute selection, learning, visualization, the experimenter that supplies the experimental environment for testing and evaluating machine learning algorithms. A simple Command-line explorer which is a simple interface for typing commands is also provided by WEKA. The limitations of this tool are weak in the documentation and classic statistics as well as poor parameter optimization [12].

### 4.8 Rapid Miner:

Rapid Miner is a software platform optimized by the same company that provides an integrated environment for machine learning, data mining and text mining predictive and business analytics. It is used for commercial and industrial applications as well as for research, education, training, rapid prototyping, application development and supports all steps of the data mining process. Rapid Miner uses the client/server model with the services provided as a program as a service or on cloud infrastructure [12].

### Feature:

• Mostly suitable for users who have experience working with a database file. this tool is platform-independent, language-independent, and supports about 22 file-formats including WEKA learning algorithm, and reading and writing an excel file [21].

• This tool is open-source software used for extracting information from the web [26].

• A Rapid Miner uses XML to describe the operator trees modeling knowledge discovery process. It has flexible operators for data input and output file formats. It contains more than 100 learning schemes for regression classification and clustering analysis[12].

### 4.9 ORANGE

ORANGE is a component-based data mining and machine learning software suite, featuring a visual programming front-end for explorative data analysis and visualization, and Python bindings and libraries for scripting. It includes a set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques [12].

### Feature

• This tool can be used as a script, work well with GUI, the easiest tool to learn and has a better debugger, it is an open-source data mining and compatible with C, C++, python, but it is weak in classical statics[21].

• ORANGE is the shortest script for doing training, cross-validation, algorithms comparison, and prediction. Because Orange is written in python hence, it is easier for most programmers to learn. ORANGE does not give optimum performance for association rules, it requires a big installation and has Limited reporting capabilities [12].

### 4.10 KNIME

KNIME is an open-source data analytics, reporting and integration platform. It has been used in pharmaceutical research but is also used in other areas like customer data analysis, business intelligence, and financial data analysis. it easily extensible [12].

### Feature:

• KNIM is a tool that integrates all analysis modules of WEKA. It is easy to learn and has the ability to interface and compatible with windows, Linux [21].

• The one aspect of KNIME that truly sets it apart from other data mining packages is its ability to interface with programs that allow for the visualization and analysis of molecular data. Some limitations in the KNIME tool are: it has only limited error measurement methods, does not has wrapper methods for descriptor selection and automatic facility for parameter optimization of machine learning/statistical methods. [12].

### 4.11 OCTOPARSE

OCTOPARSE tool mimicking human behavior to extract information and enable users to process websites that require login [8]. OCTOPARSE can help the developer to extract all the hyperlinks on the website. This gives users a simple way to automate hundreds of IPs, and at the same time provide several advanced options, like as Ajax Timeout, built-in XPath tools, etc. Besides, OCTOPARSE can crawl data for web searchers with specific requests and successfully transmit structured data [25].

## 5. Comparison table of web content mining tools

According to previous studies, it has been shown that it is difficult to compare different mining tools for web content, given the diversity of their goals and contexts, so that this paper compared these tools (as shown in Table 1) based on the following eight points[9]:

1. Open-source.
2. Platform independent.
3. Usability (User-friendly).
4. Possibility to Record the Data.
5. Perform on Structured web Data.7
6. Perform on Unstructured web Data.
7. Language.
8. Supported operating system.

Of the web content data mining tools that have been studying, it has been noted that the Screen-scrapper required prior knowledge of proxy server and some knowledge of HTML and HTTP, whereas other tools do not need any such knowledge[3], the KNIME is the tool that would be recommended for the user who is novices, Weka would be considered a very close second to KNIME because of its many built-in features that require no programming or coding knowledge because of the additional programming skills that are needed, and the limited visualization support that is provided. The Rapid Miner and ORANGE would be considered appropriate for advanced users, particularly those in the hard sciences, finally, Rapid Miner is the only tool which is independent of language limitation and has statistical and predictive analysis capabilities, So, it can be easily used and implemented on any system, moreover, it integrates maximum algorithms of other mentioned tools [12].

## 6. Conclusion and Future Work

The web data mining tools are primordial to scanning the many HTML documents, images, and text provided on Web pages. The result is provided to the search engines, in order of relevance giving more productive results of each search. The main uses of web content mining are to gather, categorize, organize and provide the best possible information available on the World Wide Web(www) to the user requesting the information. In this paper, we present a list of the available web content mining tools. Through this study, we established some objective criteria for comparison. Based on these criteria, we gave a comparative table of these different tools. From our study, all tools are examined except OCTOPARSE is undergone for study. The future scope will be predicting user needs to improve usability, scalability and user retention.

| Tool Name | Open-source | Platform independent | Record Data | Perform on structured web Data | Perform on Unstructured web Data | Usability | Language | Supported O.S |
|---|---|---|---|---|---|---|---|---|
| Web info Extractor | NO | NO | NO | YES | YES | YES | - | Windows |
| Web Content Extractor | NO | NO | NO | YES | YES | Not for unstructured data | Python | Windows |
| MOZENDA | NO | NO | NO | YES | YES | YES | Java script | Windows |
| SCRAPY | YES | YES | YES | YES | NO | YES | Python | Windows Linux, Mac, BSD |
| SCRAPER SCREEN | NO | NO | NO | YES | YES | YES | Java, Python, Jython NET,VB, ASP,PHP | Windows Linux |
| Automation Anywhere | NO | NO | YES | YES | YES | YES | Export format XML, Excel TXT, MYSQL , | Windows |
| RAPID-MINER | YES | NO | NO | YES | YES | YES | Used XML, read ,Excel file | Cross platform |
| WEKA | YES | YES | YES | YES | YES | YES | Java | Cross platform |
| ORANGE | YES | YES | YES | YES | NO | NO | Python, C, C++ | Cross platform |
| KNIME | YES | YES | YES | YES | YES | YES | JAVA | Windows Linux |

Table 1: Comparative Study of web content mining tools

## References

[1] A. Herrouz, C. Khentout, and M. Djoudi, "Overview of Web Content Mining Tools," The International Journal of Engineering And Science (IJES), vol. 2, no. 6, 2013.

[2] A.K 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). 2012 Oct;1(8). Research in Computer Engineering & Technology (IJARCET) Vol.1, no.8 , 20

[3] A. Kumar and R. K. Singh, "Web mining overview, techniques, tools and applications: A survey," International Research Journal of Engineering and Technology (IRJET), vol. 3, no. 12, pp. 1543–1547, 2016.12.

[4] A.Ranade . Halbeand A.R Joshi, "Techniques for Understanding User Usage Behavior on the Internet," International Journal of Computer Applications, vol. 92, no. 7, pp. 41–44, 2014.

[5] D. Ahamad, D. Mahmoud, and M. M. Akhtar, "Strategy and implementation of web mining tools," International Journal of Innovative Research in Advanced Engineering, vol. 4, pp. 1–7, 2017.

[6] D . M. Kene and P. K. Butey, "Web Content Mining Tools for Information Extraction in Wen Environment," International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE), pp. 141–143, 2015.

[7] D. Navadiya and R. Patel, "Web Content Mining Techniques – A Comprehensive Survey," International Journal of Engineering Research & Technology (IJERT), vol. 1, no. 10, pp. 1–6, 2012.

[8] E. Gatial, , Z. Balogh, , & Hluchý, L. (2018, June). Change Detection and Notification Method of the Rich Internet Application Content. In 2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES) (pp. 000051-000056). IEEE

[9] E. T. John, B. Skaria, and P. X. Shajan, "An Overview of Web Content Mining Tools," Bonfring International Journal of Data Mining, vol. 6, no. 1, pp. 1–3, 2016.

[10] F. Johnson and S.K. Gupta "Web content mining techniques: a survey. ," I nternational Journal of Computer Applications.vol.47,no.11 2012 .

[11] K. Pol, N. Patil, S. Patankar, and C. Das, "A Survey on Web Content Mining and Extraction of Structured and Semi-structured Data," in Proceedings of the 2008 First International Conference on Emerging Trends in Engineering and Technology, 2008, pp. 543–546.

[12] K. Rangra and K. L. Bansal, "Comparative study of data mining tools," International journal of advanced research in computer science and software engineering, vol. 4, no. 6, 2014

[13] K. Sellamy et al., "Web mining techniques and applications: Literature review and a proposal approach to improve performance of employment for young graduate in Morocco," in 2018 International Conference on Intelligent Systems and Computer Vision (ISCV), 2018, pp. 1-5

[14] M. Srividya D. Anandhi M.S. Irfan Ahmed, "WEB MINING AND ITS CATEGORIES – A SURVEY," International Journal of Engineering and Computer Science, vol. 2, no. 04 SE-Articles, Dec. 2017.

[15] Q. Zhang, and , R. S. Segall (2008 survey of current research, techniques, and software," International Journal of Information Technology & Decision Making, vol. 7, no. 04, pp. 683–720, 2008 .

[16] S. A. Inamdar and G. N. Shinde, "An agent based intelligent search engine system for web mining, "Research, Reflections and Innovations in Integrating ICT in education, 2000

[17] S. A. Taha, R.A. Shihab, and M. C. Sadik, "Studying of Educational Data Mining Techniques ,"International Journal of Advanced Research in Science, Engineering and Technology, vol. 5, no. 5, pp. 5742–5750, 2018.

[18] S. Gupta, N. Duhan, P. Bansal and J. Sidhu, "Page ranking algorithms in online digital libraries: A survey," Proceedings of 3rd International Conference on Reliability, In focom Technologies and Optimization, Noida, 2014, pp. 1-6.

[19] Sharda, D., & Chawla, S. (2012). Web Content Mining Techniques: A Study. International Journal of Innovative Research in Technology & Science.

[20] S.Huded,S. Balutagi, and A . Ranjan,"Mapping of Literature on Data Mining by J- Gate Database," in International Conference on Digital Technologies and Transformation in Academic Libraries, 2019, pp. 429–436.

[21] S. Mowla, I. Bedi, and N. P. Shetty, "A Study on Web Mining Tools and Techniques," 2017

[22] Srinath , K. R.(2017).An Overview of Web Content Mining Techniques

[23] S. Saini, and H. M Pandey,. (2015). Review on web content mining techniques. International Journal of Computer Applications,vol.118, no.18.pp.33-36,2015.

[24] S. Vijiyarani and E. Suganya, "Research issues in web mining," International Journal of Computer-Aided Technologies, vol. 2, no. 3, pp. 55–64, 2015.

[25] Thirafi, M. F. S., & Rahutomo, F. (2018, November). Implementation of Naïve Bayes Classifier Algorithm to Categorize Indonesian Song Lyrics Based on Age. In 2018 International Conference on Sustainable Information Engineering and Technology (SIET) (pp. 106-109). IEEE.

[26] T. Shanmugapriya and P. Kiruthika, "Survey on Web Content, Mining and Its Tools," International Journal of Science, Engineering and Research (IJSER) Volume, vol. 2, 2014.

[27] T. S. Kumar, M. Arthanari, and N. Shanthi, “A comparative analysis of different web content mining tools, ”International Journal of Computer, Electrical, Automation, Control and Informatoin Engineering, vol. 8, no. 9, pp. 1575–1579, 2014.

[28] V. Bharanipriya, and V. K. Prasad, “Web content mining tools: a comparative study.” International Journal of Information Technology and Knowledge Management,vol.4,no.1, pp.211-215,2011.

[29] V. Gupta and G. S. Lehal “A survey of text mining techniques and applications,” Journal of emerging technologies in web intelligence, vol. 1, no. 1, pp. 60–76, 2009.

[30] W. Fan, L. Wallace, S. Rich, and Z. Zhang, “Tapping the Power of Text Mining,” Commun. ACM, vol. 49, no. 9, pp. 76–82, 2006, doi: 10.1145/1151030.1151032.

[31] X. L. Mary, G. Silambarasan, and M. phil Scholar, “Web content mining: tool, technique & concepts,” Int. J. Eng. Sci, vol. 7, no. 5, p. 11656, 2017.