A Survey on Web Content Mining Techniques and Its Tools

Yi Yi Aung, Ei Ei Thu
University of Computer Studies, Mandalay
yiyiaung123@gmail.com, eieithuet@gmail.com

Abstract

The World Wide Web (WWW) is a rich source of information and continues to expand in size. The complexity at the WWW makes it difficult for data mining. Before the Web, finding information meant asking a friend or an expert, or buying/borrowing a book. However, it is becoming a challenging task to effectively and efficiently retrieve the required web page /information on dynamic and heterogeneous web documents. If we can efficiently use web content mining techniques and tools, we can easily extract useful information from various forms of web page content by eliminating noisy information. This paper presents different web content mining techniques such as structured, unstructured, semi-structured and multimedia and five tools that can collect the appropriate information from websites that the user requires. The focus of this paper is to bring to light the value of Web Content Mining techniques and its tools in virtual society.

Keywords: *Web content mining, Web content mining techniques*

1. Introduction

The World Wide Web is a collection of inter-related files on one or more web servers. Most people browse the internet for retrieving information, but its size and complexity is increasing day by day. People must use more time as they access lots of irrelevant documents.

In effectively and efficiently retrieving information from the web, web mining techniques and tools are used. The term web data mining is a technique used to crawl through various web resources to gather required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. There is an increasing trend among companies, organizations and individuals to gather information through web data mining to utilize that information in their best interest. It is the data mining technique that automatically discovers/extracts the information from web documents. [8] It is a task to search relevant information from huge amount of data. Web mining tasks can be categorized into three types: web structure mining, web usage mining and web content mining.

Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. [5] Web usage mining is the process by which we identify the browsing patterns by analyzing the navigational behavior of user. It uses the secondary data on the web. [8]

Web content mining is mining data from the content of web pages. Web pages consist of text, graphics, tables, data blocks, data records and multimedia data. Web content mining uses the ideas and principles of data mining and knowledge discovery process.

2. Related Work

Current Internet includes billions of pages consist of various kind of data layout. Whether to convert existing sites or sites semantics for intelligent use embedded data, the definition of data mining techniques and tools are of great interest in research.

Abdelhakim Herrouz, Chabane Khentout and Mahieddine Djoudi [1] demonstrated that the mining tools are imperative to scanning the many HTML documents, images, and text.

Arvind Kumar Sharma and Gupta P.C.Gupta [2] presents to increase such a process the software related to web content mining can be used so that a computer can use this software or tools to download the essential information that one would require.

Faustina Johnson and Santosh Kumar Gupta [4] explain the paper by using four sections. Firstly, this paper introduces web mining, secondly, it tries to explain the interrelationship of web mining with various other areas, thirdly, it explains various web content mining techniques and finally the paper concludes with the analysis of these various techniques.

T.Shanmugapriya and P.Kiruthika [9] discuss the knowledge that web content mining becomes complicated when it has to mine structured, semi-structured, unstructured and multimedia data. The different types of web content mining tools and its features are also discussed.

3. Web Content Mining

Web content mining utilizes the automatic way of information extraction from the World Wide Web according to the preferences. It is related to data mining and text mining because much of the web contents are text based. But all of its techniques are not based on them. Web content mining is the mining, extraction and integration of useful data, information and knowledge from web page content. However, data in its broader form has to be further narrowed down to useful information. [4]

The web content data can be unstructured form such as free text or in structured form such as data in the tables or in semi structured form such as HTML documents or multimedia data. Different techniques and tools are required to produce a higher quality of information to the user. [4] Research in web content mining encompasses resource discovery from the web, document categorization and clustering, and information extraction from web pages. [7]

Web content mining could be separated from two points of view: the agent-based approach or the database approach. The three types of agents are intelligent search agents, information filtering/categorizing agent and personalized web agents. The first approach aims on improving the information searching and filtering. The second approach aims on modeling the data into more structured form in order to apply standard database querying mechanism. The approach can contains well-formed databases containing schemas and attributes with defined domains. Technologies used in web content mining are NLP, IR. [4][10]

4. Web Content Mining Techniques

The concept of "WEB CONTENT MINING" involves techniques for summarizing, classification and clustering of the web contents. It can provide useful and interesting patterns about user needs and contribution behavior. [7]

Some of the prominent web content mining techniques are as follows: [9]

1. Unstructured data mining techniques
2. Structured data mining techniques
3. Semi structured data mining techniques
4. Multimedia data mining techniques

4.1. Unstructured Data Mining Techniques

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. [2] Some useful unstructured data mining techniques are described as follows. The input, output and method of unstructured data mining techniques are described in table 1.

4.1.1. Information Extraction

Pattern matching is used to extract information from unstructured data and convert into structured data. Pattern matching and transformation are also used. [2]

4.1.2. Topic Tracking

In topic tracking, it tracks the topics searched by the user and predicts the documents and produce to the user that of interest. Prediction techniques are used. [2]

4.1.3. Categorization

Categorization is the technique of categorizing the document. [4] Categorization can be used in business and industries to provide customer support. [2]

4.1.4. Clustering

It is very difficult to find out the relevant information from large unstructured document collection. We categorize the documents using categorization technique. Same document can appear in different groups. The problem of finding best such grouping can be handled by clustering. [4]

4.1.5. Summarization

Two methods used in summarization are extractive and abstractive. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. Abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary. [4]

Table 1. Unstructured Data Mining Techniques

Technique	Input	Method	Output
Information Extraction	Document	Pattern matching and transformation	Structured data
Topic Tracking	Database	Topic filtering from database, Topic Re-ranking Topic tracking	Related topics

Categorization	Document	Count the words Explore main theme Ranking	Related document according to ranking
Clustering	Document Collection	Categorization Exploring best group	Best relevant group containing desired document
Summarization	Document	Extractive Method Abstractive Method	Short set of words that conveys the message of the document

4.2. Structured Data Mining Techniques

The techniques used in mining structured data are presented as follows. The analysis of structured data mining techniques is described in table 2.

4.2.1. Intelligent Web Spiders

Web spiders or crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. It can also be used for automating maintenance tasks on a Web site such as checking links or validating HTML code. [4]

4.2.2. Wrapper Generation

Wrapper is a set of extraction rules to extract the data from the web pages, this can done either manually or automatically. Meta Search Engines are connected to search engines by the means of Wrappers. [4]

Table 2. Structured Data Mining Techniques

Techniques	Web Spiders	Wrappers
Method	Crawls through hyperlinks. Creates Index Database	Connect various search engines. Passes the query from one search engine to another. Resolves formatting issues existing between Search Engines
Use	Search Engines Textual Analysis Access Market Trends	Meta Search Engines
Example	Google Yahoo	Visual Web Ripper Screen-Scraper

4.3. Semi-Structured Data Mining Techniques

Semi-structured data arises when the source or environment does not impose a rigid structure on the data when data is combined from several heterogeneous sources. [4] This paper presents some of useful Semi-structured data mining techniques. The analysis of semi-structured data mining techniques is described in table 3.

4.3.1. Object Exchange Model

A main feature of object exchange model is self describing; there is no need to describe in advance the structure of an object. [2]

4.3.2. 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. The central idea of this approach is to find out objects identical to the object we are considering. [4]

Table 3. Semi-Structured Data Mining Techniques

Technique	Object Exchange Model	Top Down Extraction
Method	Connect various search engines. Passes the query from one search engine to another. Resolves formatting issues existing between Search Engines	Extract Complex Objects Decompose complex objects Extract atomic objects
Advantage/ Disadvantage	Simple Effective	Requires previous knowledge of data source. Additional work required when source changes.

4.4. Multimedia Data Mining Techniques

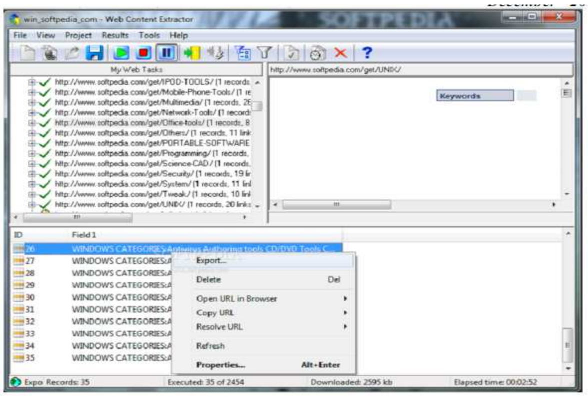
Multimedia data mining can be used as the process of finding interesting patterns from media data such as audio, video and image that are not ordinarily accessible by basic queries and associated results. Audio data contains radio, speech or spoken language. To mine audio data one could first convert it into text using speech transcription techniques and then mine the text data. Video mining involves finding association between video clips and to find out unusual pattern in video clips. Image processing focuses on detecting abnormal patterns as well as retrieving images. Image mining is all about finding unusual patterns. [4]

5. Web Content Mining Tools

This paper presents a variety of web content mining tools. [10] There are various tools available to produce user interested knowledge by using various web mining techniques as shown in table 4 and the comparison of tools can be seen in table 5.

5.1. Web Content Extractor (WCE)

Web content extractor is the most powerful data extraction tool designed for web scraping, data mining, and data extraction without any programming. The purpose of web content extractor is to simulate the human browsing experience by replicating computer commands to click, and ultimately extract specific data fields from a web page. [10] The scraped data of web content mining according to rules and patterns can be seen in figure 1.



http://www.softpedia.com/get/UNIX/

- http://www.softpedia.com/get/IPOD-TOOLS/ (1 records
- http://www.softpedia.com/get/Mobile-Phone-Tools/ (1 re
- http://www.softpedia.com/get/Multimedia/ (1 records, 26
- http://www.softpedia.com/get/Network-Tools/ (1 record
- http://www.softpedia.com/get/Office-tools/ (1 records, 8
- http://www.softpedia.com/get/Others/ (1 records, 11 link
- http://www.softpedia.com/get/PORTABLE-SOFTWARE
- http://www.softpedia.com/get/Programming/ (1 records,
- http://www.softpedia.com/get/Science-CAD/ (1 records,
- http://www.softpedia.com/get/Security/ (1 records, 19 li
- http://www.softpedia.com/get/System/ (1 records, 11 lin
- http://www.softpedia.com/get/Tweak/ (1 records, 10 lin
- http://www.softpedia.com/get/UNIX/ (1 records, 20 links

ID	Field 1
26	WINDOWS CATEGORIES:Antivirus Authoring tools CD/DVD Tools C...
27	WINDOWS CATEGORIES:A
28	WINDOWS CATEGORIES:A
29	WINDOWS CATEGORIES:A
30	WINDOWS CATEGORIES:A
31	WINDOWS CATEGORIES:A
32	WINDOWS CATEGORIES:A
33	WINDOWS CATEGORIES:A
34	WINDOWS CATEGORIES:A
35	WINDOWS CATEGORIES:A

Expor Records: 35 Executed: 35 of 2454 Downloaded: 2595 kb Elapsed time: 00:02:52

Figure 1. Web Content Extractor

The features of web content extractor are as follow. [11]

- This tool helps businessmen extract and collect the market figures, product pricing data or real estate data.
- It is a template of web data extraction tool to extract specific data from a website automatically.
- It is a customized web crawler / web spider.
- This tool assists Journalists to extract news and articles from news sites. [2],[10]
- It uploads the output file to a FTP server.
- It extracts data from password protected websites.
- It is easy to use configuration wizard and very simple to use, quick learning curve and right to the point. [11]

5.2. Web Info Extractor (WIE)

Web Information Extractor is helpful in mining web relevant document from collections and monitoring content update. It can extract structured or unstructured data from web page, save to database, reform into local file or post to web server. No need to define complex template rules, just browse to the web page you are interesting and click what you want to define the extraction task, and run it as you want, or let it run automatically.

The features of web info extractor are as follow. [2], [10], [12]

- It is Easy to define extraction task and no need to learn complex and boring template rules.
- Extract unstructured data as well as tabular data to file, database.
- Monitor web pages and extract new content when update.
- It can deal with text, image and other link file.
- Unicode support can process web page in all languages.
- It can run multiple tasks at the same time.
- Support recursive task definition.

5.3. Automation Anywhere (AA)

Automation anywhere is a powerful and easy web data extraction tool used for retrieving web data effortlessly, screen scrape from web pages are use it for web mining. The intelligent automation software, used for automating and scheduling business process and IT tasks provides easier way. [11] It can provide automate and schedule complex tasks in minutes, without any programming. Record keyboard and mouse or create automate scripts with drag and drop actions. [2] The features of automation anywhere are as follow. [10][3]

- It is an advanced pattern matching and extraction technology.
- Intelligent automation is used for business and IT tasks.
- Web recorder with advanced data extraction technology plug into internet explorer and enables you to extract and validate data-reliably and with ease.
- Creating automation tasks takes few minutes, record keyboard and mouse stokes, or uses easy point-and-click wizards.
- Integration with excel, database, or any application which lets you store the extracted web data into any application in any format of your choice.

5.4. Screen Scraper

Screen-scraping is software that retrieves or “scrapes” information off a variety of websites simultaneously. Like a database, it permits to mine the data of the World Wide Web. It permits mining the content form the web, like searching a database, SQL server or database, which interfaces with the software, to achieve the content mining requirements. [2] [13] The process of screen scraper are described in figure 2.

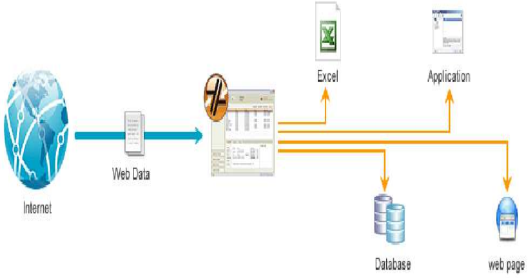


Figure 2. Screen Scraper

The features of screen-scraper are as follow.

- Graphical interface is provided by the Screen-scraper allowing you to designate URL's, data elements to be extracted and scripting logic to traverse pages and work with mined data.
- The programming languages like Java, .NET, PHP, and Active Server Pages can also be used to access screen scraper.
- One of the most regular usages of this software and services is to mine data on products and download them to a spreadsheet. [10]

5.5. Mozenda

Mozenda is a software tool that enables anyone to become a web archaeologist, digging and extracting information from the web. It can support commercial and nontechnical users to simply mine data across web pages. For those who know what they are looking for and the value of the information they receive, Mozenda is a priceless tool in their technology backpack. Once information is in the Mozenda systems users can reuse, format, and mash up the data to be used in other online/offline applications or as intelligence. [2] The two part of Mozenda's scraper software which are building data extraction project called agent in window application and exporting extracted data in cloud are described in figure 3. [6]

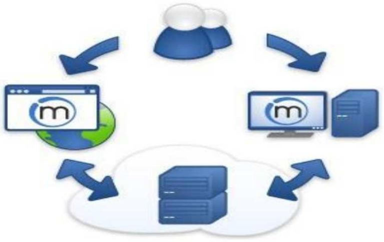


Figure 3. Mozenda

The features of mozenda are as follow.

- Mozenda can load pages and navigate pages just like a browser and can easily navigate through sub-pages.
- Mined data can be accessed online, exported, as well as used throughout an API.
- Mozenda data extractor is an excellent tool that performs user scraper within the clouds. [10], [11]

Table 4. Various Web Content Mining Tools

No	Tool	Description according to usage
1.	WCE	It is a tool that uses data extraction for web scraping, data mining or data extraction from the web page.
2.	WIE	It helps to extract structured or unstructured data from Internet, reform into local file or save to database, place into web server.
3.	AA	Use Automation anywhere to automate scripts in disparate formats.
4.	Screen Scraper	It is a tool for extracting information from web sites. It can be used for searching a database, SQL server or SQL database.
5.	Mozenda	It helps to enables users to extract and manage web data. Users can setup agents that routinely extract, store, and publish data to multiple destinations.

Evaluation Criteria: The different Web Content Mining tools are complex and difficult to compare because of the diversity of goals and contexts. We compare the existing tools depend on the following four facts: [1]

- Usability (User friendly)
- Possibility to Record the Data
- Perform on Structured web Data
- Perform on Unstructured web Data

Comparative Table: The following table shows the summary of the characteristics of Content Mining Tools. In the table, we use the following symbols: [1]

- - :For No
- ✓ : For Yes
- U : Usability
- R.D : Record the Data
- E.S.D : Extract Structured web Data
- E.U.D : Extract Unstructured web Data

Table 5. Comparison of Web Mining Tools

Tool	U	R.D	E.S.D	E.U.D
WCE	-	-	✓	✓
WIE	✓	-	✓	✓
AA	✓	✓	✓	✓
Screen-Scraper	-	-	✓	✓
Mozenda	✓	-	✓	✓

5.6. Commonalities and Differences between the above tools

Some commonalities and differences between web content mining tools are described as follows:

1. Commonalities

- All the tools automate the business task and retrieve the web data in an efficient way.
- All the tools are performed on structured and unstructured web data.

2. Differences

- AA5.5 allows recording of action. This facility unique and it is not provided in the other tools.
- Screen-scrapper needs prior understanding of proxy server and some awareness of HTML and HTTP and it need internet connection to run.
- Though we have setup file, Mozenda will not allow us to install without Internet connection, this is not the case with other tools.

6. Conclusion

The World Wide Web is the universe of network-accessible information that increases in size and complexity with time hence making it difficult to extract relevant information. Searching a topic from web is difficult to get accurate topic information but now a day it is easy to get the proper and relevant information due to web mining. In web mining, we can use not only techniques or algorithms but also tools to get useful information and knowledge from web page contents. We have to apply different kinds of techniques depend on different types of data structure. Each techniques and tools has their respective features. These can be used on different application areas depend on the system needs. The way of searching information also differs in respect to people knowledge, the content of the page and necessity of them. By choosing appropriate techniques and tools in our system we can make our search of contents over the web faster and exact. The paper discusses some of the nature of web content mining techniques, tools and their respective features to extract useful information and knowledge from web page contents.

References

[1] Abdelhakim Herrouz, Chabane Khentout and Mahieddine Djoudi,"Overview of Web Content Ming Tools", International Journal of Scientific Engineering and Research (IJSER), ISSN: 2319-1813 ISBN: 2319-1805, Volume 2, Issue 6,2013.

[2] Arvind Kumar Sharma and Gupta P.C.Gupta, "Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining", International Jouranal of Advanced Research in Computer Engineering & Technology (IJARCET), ISSN: 2278-1323, Volume 1, Issue 8, October 2012.

[3] Automation Anywhere 5.5 help

[4] Deepti Sharda and Sonal Chawla, "WEB CONTENT MINING TECHNIQUES: A STUDY", International Journal of Innovative Research in Technology & Science (IJIRTS), ISSN:2321-1156.

[5] http://en.wikipedia.org/wiki/Data_mining.

[6] http://scraping.pro/mozenda-review/

[7] Kanika Dhingra and Govind Murari Upadhyay, "Web Content Mining: Its Techniques and Uses", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 3, Issue 11, November 2013.

[8] K injal Patel and Niki R.Kapadia, "Web Content Mining Techniques – A Comprehensive Survey", IJREAS, ISSN: 2249-3905, Volume 2, Issue 2, and February 2012.

[9] P. Kiruthika and T. Shanmugapriya, "Survey on Web Content Mining and Its Tools" International Journal of Scientific Engineering and Research (IJSER),ISSN (online): 2347-3878, Volume 2 Issue 8, August 2014.

[10] P.Ponmuthuramalingam and S.Balan, "A Study of Various Techniques of Web Content Mining Research Issues and Tools", International Journal of Innovative Research & Studies, ISSN 2319-9725, Volume 2, Issue 5, May 2013.

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

[12] Web Info Extractor, http://webinfoextractor.com/

[13] www.screen-scraper.com