Al-Qadisiyah Journal of Pure Science

Volume 29 | Number 2 Article 1

12-20-2024

A comprehensive study of content-based web mining

Noor Al-Deen Alaa Mohammed Tahaa
University of Technology - Computer Science Department / Baghdad, nooraldeen072@gmail.com

Israa Tahseen Ali
University of Technology - Computer Science Department / Baghdad, israa.t.ali@uotechnology.edu.iq

Recommended Citation

Tahaa, Noor Al-Deen Alaa Mohammed and Ali, Israa Tahseen (2024) "A comprehensive study of content-based web mining," *Al-Qadisiyah Journal of Pure Science*: Vol. 29 : No. 2 , Article 1.
Available at: https://doi.org/10.29350/2411-3514.1278

ORIGINAL STUDY

A Comprehensive Study of Content-based Web Mining

Noor Al-Deen A. Mohammed Tahaa*, Israa T. Ali

University of Technology, Computer Science Department, Baghdad, Iraq

Abstract

With the huge growth of the web, and with the research in information retrieval. For many companies, websites have become the primary medium of communication and information. It also provides the potential for data mining and unprecedented opportunities and challenges. There are many different techniques used to extract useful information from the web. This information is extracted to assist users in better understanding the structure of content on the internet. In this research paper, web mining is described in a simplified manner and its types, with a detailed focus on web content mining and its main approaches that are used to mine the data with the latest techniques used in this field.

Keywords: Data mining, Web mining, Web usage mining, Web content mining, Web structure mining

1. Introduction

The World Wide Web is a massive source of information. Its complexity and size are constantly increasing. Many challenges start to appear in retrieving the needed web pages in the www, effectively and efficiently. When the needed pages are searched for by a user, he or she wants those relevant pages to be at hand. The relevant information becomes very difficult to extract, filter, evaluate or find for the users, because of the huge amount of information. So, the need for techniques that solve these challenges becomes very important. With the help of some areas like machine learning, database (DB), natural language processing (NLP) and information retrieval (IR), etc. the executing of web mining can be easily done [1].

Web mining can be executed on semi-structured or unstructured data like texts and it extracts information from web services and documents automatically. Many types of data, web mining focus on, like user access information, contents of web pages, hyperlinks between pages, and many web resources to have in hand the best properties among the data objects [2].

The following issues are generally mentioned in research and applications that are associated with the web [3]:

- Finding relevant information [3].
- Finding needed information [3].
- Learning useful knowledge (Web mining) [3].
- Personalization/recommendation of information [3].

2. Uses of web data mining

There are many benefits areas of web mining, some of these benefits are briefly discussed as below [2]:

2.1. E-learning

Web mining utilized to improve and enhance e-learning environments' methods. Web mining for e-learning is primarily based on web usage, i.e., online rather than offline. Working on improving web-based learning environments is one of the main duties performed by machine learning technology and online usage mining approaches [2].

Received 4 February 2022; accepted 29 June 2022.
Available online 18 April 2025

* Corresponding author.
E-mail addresses: nooraldeen072@gmail.com (N. Al-Deen A. Mohammed Tahaa), israa.t.ali@uotechnology.edu.iq (I.T. Ali).

https://doi.org/10.29350/2411-3514.1278

2.2. Libraries on the web

Digital library services provide valuable distributed data all over the world, decreasing the need to physically visit different libraries in different parts of the world [2].

2.3. E-commerce

One of the most difficult aspects of e-commerce is gaining a thorough understanding of visitors' or customers' demands and price preferences. It increase the service's capacity for clients and provide competitive advantages [2].

2.4. Electronic government

Organizations that work with the country's population provide better social services. The greatest distinguishing feature of e-government systems is the use of technology to offer services electronically, focusing on citizens' needs by delivering more information and services in support of the government. E-government systems may provide citizens with custom-tailored services, resulting in increased user satisfaction, service quality, and citizen decision-making support, all of which result in social benefits [2].

2.5. E-democracy and e-polities

E-Polities provides political information and 'politics on demand' to citizens improving political transparency and democracy. Election information, parties, members of parliament, and members of native governments on the internet are a part of e-politics services. Despite the importance of e-politics in democracy, there are restricted web mining strategies to fulfill citizens' desires [2].

2.6. Electronic business

Web mining strategies will help a web-enabled electronic business improve marketing, customer service, and sales [2].

2.7. Crime investigation and security investigation

Fraudulent websites, hacking, porn distribution, misappropriated online gambling, internet fraud, virus propagation, and cyber-terrorism are all examples of cyber-crime. Web mining techniques are used to safeguard user systems of this logging information. Web mining techniques like classification and clustering will expose cybercriminals' identities, whereas decision trees, neural networks, genetic algorithms, and support vector machines are frequently used to trace crime patterns and visualize networks on websites [2].

3. Web mining techniques

Web data mining is the application of data mining, and it has three techniques, the first one web content mining used to mine or extract knowledge or useful information from web pages, the second is web structure mining used to find useful information and knowledge from hyperlinks structure, the third is web usage mining used to discover the access patterns of the user from web usage logs [3]. Fig. 1 shows the web mining techniques [4]:

3.1. Structure-based web mining

Web structure mining is also called, "Link analysis.". WSM is an old area of research, and the interest in this research has been increased in the latest days, because of the increasing interest in web mining. So, a new research area called link mining has appeared [5].

3.2. Usage-based web mining

Web usage mining is outlined as the detection and analysis of patterns automatically in user transactions, clickstreams, and different data that are generated or collected as a result of users' interaction with internet resources on websites. It's a way of understanding user behavior on the web [6].

3.3. Content-based web mining

Web content mining is used to mine, extract and scan the contents of web pages like text, graphs, videos, and images. There are two approaches that are used with (WCM). The first one is the database approach that is used to help to retrieve from web documents the data that is semi-structured. The second is an agent-based approach that searches at the information that is relevant and organizes it. As for the data available on the web, most of them are unstructured data. There are two viewpoints of (WCM) and they are the retrieval of information view and database view. The primary goal of (WCM) from the first view (information retrieval) is to enhance the finding and filter information to the clients, and also the managing of the data web is the task of the database (DB) view [7]. Fig. 2 shows the web content mining techniques. [8], where WCM

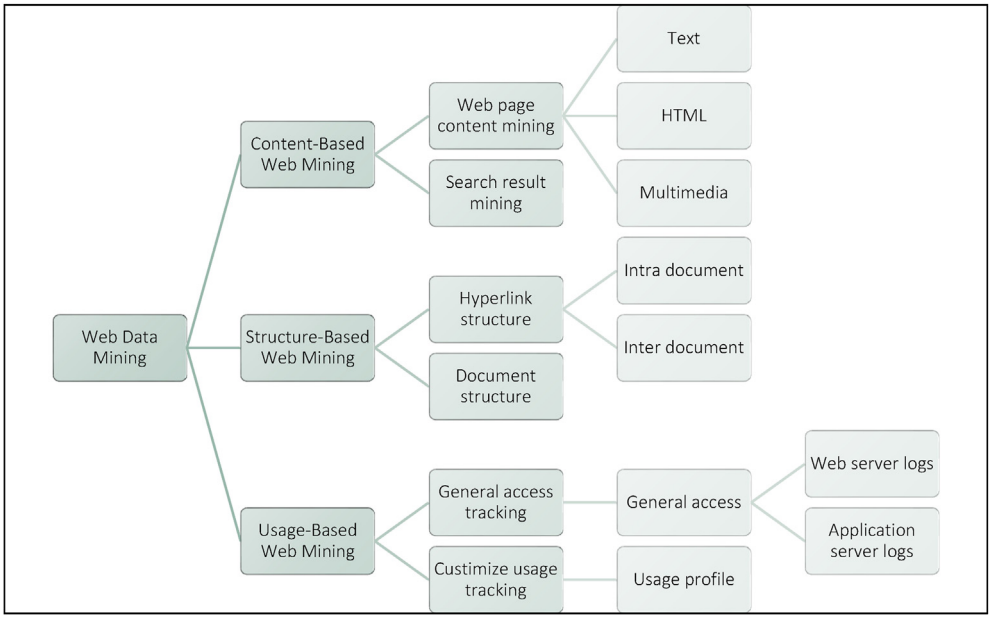

Fig. 1. Web mining techniques [4].

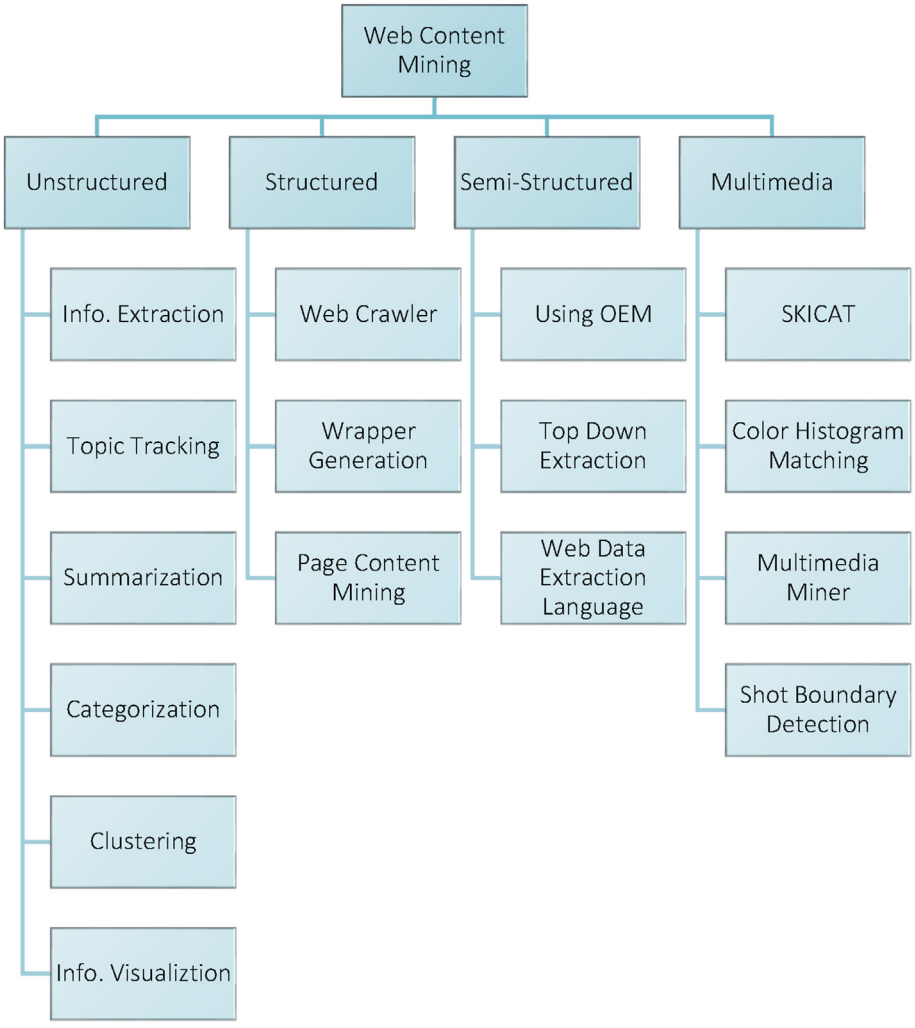

Fig. 2. Web content mining techniques [8].

has several approaches used to mine the data, and they are as follows [7]:

3.3.1. Unstructured data mining

The available web data are mostly unstructured, and text documents are unstructured. The process of applying DM techniques to the data that are unstructured is called knowledge discovery in texts. So, the used techniques for this process are summarization, information extraction, clustering, topic tracking, categorization, and information visualization, and they are as follows [7]:

- **Extraction of information technique**

This technique is using pattern matching to extract information from data that are unstructured on the web. It finds the connection of keywords within the text by tracing the phrases, and words. This technique is very useful when the size of the text is large on the web. The extraction information converts the text to a more structured form. The first step is to mine the information from the extracted data, and the second step is to find the missed information by using different types of rules. Information extraction discards the incorrect prediction [7].

- **Tracking of topics technique**

This method examines the documents that the client views and analyzes the user profile. This technique is used to predict the document that is connected to the user's interest. Therefore, yahoo has applied this technique, when the user enters a keyword then anything is connected or related to this keyword, the user will be informed about it. Topic tracking is mostly used in two fields, the medical field, and the education field. The disadvantage of topic tracking is that it may give the user material that is unrelated to his or her topic when he or she searches for it [7].

- **Summarization technique**

Summarization is used to cut down on the length of the document while keeping the most significant points. It assists the user in determining whether or not the document is important. This technique can be completed in a very short amount of time. The first way is the extractive method, which involves picking a subset of words, phrases, and sentences from the source text to construct the summary. The second method is the abstractive method, which involves creating an internal semantic representation before constructing the summary using the natural language (NL) generating methodology. This summary may include words that are not included in the original source [7].

- **Categorization technique**

Categorization determines the primary topic by grouping documents into predetermined categories and counting the number of words in each document. The primary topic is then determined. As a result, the document is given a rating based on the topic. The highest priority is given to documents with the best topic. Customer service is provided by categorization for businesses and industries [7].

- **Clustering technique**

Similar documents are grouped together using the clustering technique [7]. Clustering is done based on the fly, not on a predefined topic. That is a different group may have the same document that is in another group. Therefore, important documents won't be omitted from the search results. This technique helps a user to pick out an interesting topic. Clustering is incredibly helpful in management information systems [9].

- **Visualization of information technique**

Feature extraction and key term indexing are used in this method. The documents' commonalities are discovered through visualization. As a result, vast textual materials are displayed as hierarchies or visual maps wherever browsing is permitted. It aids in the visual analysis of content. The client will be able to interact with the graphs by zooming, scaling, and creating sub-maps [7].

3.3.2. Structured data mining

For mining structured data three techniques are used, the first is a web crawler, the second is wrapper generation and the third is page content mining, and they are as follows [9]:

- **Web crawler technique**

Web crawlers are computer programs and there are two types of them called external and internal web crawlers. Web crawlers' task is traversing the structure of the hypertext on the web. The web crawler that is external crawls through the unknown website, and the web crawler that is internal crawls through internal pages of the internet site that is returned by the crawler that is external [9].

- **Wrapper generation technique**

Wrapper generation is a technique that offers information about the potential of sources. Traditional search engines have already given websites a grade. The value of page rank is used to obtain web pages in accordance with the query. The sources determine what query they'll respond to, and so the output varies. The wrappers can also provide meta-information about the sources, such as domains, statistics, and index operations [9].

- **Content mining of a page technique**

This technique is a structured data extraction strategy that works on standard search engine-ranked pages. As a result, this technique classifies the pages by comparing page content rank [9].

3.3.3. Semi-structured data mining

Top-Down extraction, object exchange model (OEM), and web data extraction language are the techniques that are used for semi-structured DM and they are as follows [9]:

- **Object exchange model (OEM) technique**

The important information is retrieved from semi-structured data and incorporated in a group of useful information, which is then saved in a database (OEM). It aids the user in better comprehending the internet's information structure. It's ideal for a dynamic, heterogeneous environment. A key property of the OEM is that it is self-descriptive; there is no need to explain the object structure beforehand [9].

- **Top-down extraction technique**

From a set of rich web sources, this technique extracts complex objects and changes them into less complicated objects until the extracting is done of the atomic objects [9].

- **Web data extraction language technique**

Converting the data of the web to structure data and delivering it to end-users is done by web data extraction language. And in the form of tables, the data is stored by this technique [9].

3.3.4. Multimedia data mining

Color histogram matching, multimedia miner, SKICAT, and shot boundary detection are examples of multimedia data mining methods, and they are as follows [8]:

- **SKICAT technique**

This technique can be the next big thing in astronomical data analysis and cataloging, as it creates a computerized catalog of sky objects. It converts these objects into human-readable classes using machine learning approaches. It combines image processing and data classification techniques to help classify a big classification set [8].

- **Color histogram matching technique**

This technique consists of firstly, equalization of the color histogram that is trying to find the correlation between color components, and secondly smoothing which is used to solve the problem of the presence of unwanted artifacts in the equalized image [7].

- **Multimedia miner technique**

There are four major steps during this technique, the primary is image excavator which is the method of extracting video and image, the second is a preprocessor to extract features from the image and these features are held on in a database, the third component is a search kernel, which is used to match queries with the database's video and image resources. Image information mining techniques are used by the discovery module to find out patterns in photos [8].

- **Shot boundary detection technique**

To recognize video boundary shots automatically, shot boundary detection is used [8].

4. Conclusion

Web mining is a rapidly expanding field of study. Web content mining is similar to text mining and data mining, but they are not the same. Web content mining deals with data that is either unstructured or semi-structured. There are several approaches to web content mining, and each approach uses a different set of techniques to mine the data. Unstructured data mining, structured data mining, semi-structured data mining, and multimedia data mining are the web content mining methodologies that are used to mine the data.

Funding

This research received no specific grant from any funding agency in the public or commercial, and it is self-funding.

References

[1] Bhatia Tamanna. Link analysis algorithms for web mining. Int J Comput Sci Telecommun 2011;2(2):243–6. 4333.

[2] yadav S, ahmad K, shekar J. Analysis of web mining applications and beneficial areas. IIUM Eng J 2011;12(2):185–95.
[3] Raju Y, Suresh Babu D. A novel approaches in web mining techniques in case of web personalization. Int J Res Comput Appl Robot 2015;3(2):6–12.
[4] Vijiyarani S, Suganya E. Research issues in web mining. Int J Comput Appl Technol 2015;2(3). https://doi.org/10.5121/ijcax.2015.2305.
[5] Chopra Preeti, Ataullah Md. A survey on improving the efficiency of different web structure mining algorithms. Int J Eng Adv Technol 2013;2(3):296–8.
[6] Hulliyyatus Suadaa Lya. A survey on web usage mining techniques and applications. In: 2014 international conference on information technology systems and innovation, IEEE, ICITSI 2014 - proceedings; 2014. p. 39–43. https://doi.org/10.1109/ICITSI.2014.7048235.
[7] Saini Shipra, Pandey Hari Mohan. Review on web content mining techniques. Int J Comput Appl 2015;118(18):33–6. https://doi.org/10.5120/20848-3536.
[8] Kumar Sharma Arvind, Gupta PC. Study & analysis of web content mining tools to improve techniques of web data mining. Intern J Adv Res Comput Eng Technol (IJARCET) 2012;1(8):287–93.
[9] Johnson Faustina, Kumar Gupta Santosh. Web content mining techniques: a survey. Int J Comput Appl 2012;47(11):44–50. https://doi.org/10.5120/7236-0266.