Preprints.org

Review

Not peer-reviewed version

Web Content Mining: A Review on Concepts, Techniques, and Tools

Ali Hassan Sial *

Posted Date: 29 July 2024

doi: 10.20944/preprints202407.2339.v1

Keywords: WWW; Web Content Mining; Unstructured; Structured; Semi-Structured; Web Documents

Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2024 doi:10.20944/preprints202407.2339.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Review

Web Content Mining: A Review on Concepts, Techniques, and Tools

Ali Hassan Sial [1], Muhammad Ayoub Kamal [1] and Kamlesh Kumar [2]

[1] Department of Computer Science, DHA Suffa University (DSU), Karachi, Pakistan; ayoub.kamal@dsu.edu.pk
[2] Department of Software Engineering, Sindh Madressatul Islam University (SMIU), Karachi, Pakistan; kamlesh.kumar@smiu.edu.pk
* Correspondence: hassan.sial@dsu.edu.pk

Abstract: With the emergence of the Internet and WWW has become a comprehensive source of information for the extraction of meaningful information has become a significant challenge in the past decade. Furthermore, the available information is classified in unstructured, structured, and semi-structured forms. Numerous amount of information on the web is also characterized in an unstructured or semi-structured format. This usually extracts the potential useful information from these multiple forms and has been considered a leading area of research in this modern era. The authors presented web content mining as a subcategory of web mining that focuses on important extracting patterns from the contexts available in the documents on the web. Hence, this paper focuses on the multiple content mining concepts, tools, and techniques implemented on the documents available on WWW.

Keywords: WWW; Web Content Mining; Unstructured; Structured; Semi-Structured; Web Documents

1. Introduction

The rapid advancement in technology has led to the emergence of the World Wide Web. In this digital era, WWW has evolved in each aspect of human life. It has expanded the quantity of information based on the customers' expectations regarding the usage and performance measurement aspects [1]. This notion needs regular updates, priorities, and strategies to fulfill the users' basic essentials and site visitors.

Furthermore, big data and data mining principles are rapidly increasing in every aspect of a real-world scenario. These techniques are highly delving into the World Wide Web. Generally, the term WWW refers to collecting documents, textual data, files, images, and other relevant mediums of data in structured, unstructured, and semi-structured forms comprising a higher level of diversity, accuracy, and dynamic modular framework for increasing the rate of scalability and minimizing the redundancy rate.

The key approach of web mining is to extract relevant and valuable facts and figures from the WWW. In this case, web data stores' evolution is considered the essential source of information for several users in multiple domains. Although, web mining is usually a challenging aspect based on the identities and lack of structure in web services. Based on similar conditions, the internet users recently covered the facts and figures on contrary to the excess of knowledge, data or information [2]. Several web users would possibly encounter the following issues, however the collaboration with the web or internet is possible with the following elements:

© 2024 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2024 doi:10.20944/preprints202407.2339.v1

1.1. Searching Appropriate Information

Whenever a user searches for the relevant facts and figures in the WWW, they contribute a simple keyword in the query. The query response will be on the ranking of page lists based on the relevancy to the search query. Likewise, nowadays, many search engine tools constitutes of specific issues, i.e. less precision rate. It usually occurs because to the irrelevant information of search results and less redundancy rate (usually occurs in terms of inability to index all the available information).

1.2. Knowledge Discovery from WWW

Knowledge Discovery is a data-triggered process, however the initial step is a query-based procedure. In this phase, the user working on the WWW has to extract precise and valuable information from a wide variety of acquisition-driven contextual frameworks.

1.3. Data Personalization

Data Personalization is associated to the category and production of statistics, as people usually distinguish in terms of contextual data and presentations they usually prefer when interacting with each other.

1.4. Analysis of Individual User Preferences

This feature encounters a problem that is an essential requirement of all site users. This comprises of the personalization of distinct users, website design and administration, customizing user information, etc. The web develops loud as if it consists of various types of information. Hence, the web mining strategies can be utilize to solve these concerns.

2. Related Work

Dr. John, Eldhose, T. and *et al.,* presented an overview on web content mining tools and techniques, the sole purpose of their research work is to examine and explain the conceptual framework, web mining platforms for exchanging information, as it is easier to publish their documents in an efficient and effective manner. The increase in number of users and service providers' increases, the number of documents grows and search patterns of information are becoming more complicated and time-consuming process. The authors have provided concepts associated with data mining and web mining along with their tools, and comparative tables of such tools as per their relevant criteria [1]. Satish, N. R., conducted a study on multiple applications, approaches and problems associated with Web Content Mining, the internet has drastically become the most prevalent essence of information in this technological era, as the rapid expansion of web increases, enormous amount of data and diverse kinds of information are storing online at the fast paced. The authors have presented various surveys and examining of web content mining methods and applications [2]. Yu, Zhaohan wrote a thesis on optimization techniques in the data mining with applications as per biomedical and psychophysical datasets, the author discussed about various techniques, concepts, algorithms, and mining tasks to analyze and interpret large amount of datasets and relevant sources of information [3]. Samuel MakindeOpeyemi, and *et al.,* presented a review on the contemporary trends in web content mining, the understanding and examining of web documents, the relevancy of webpages and several other areas are used as developing areas in web mining. Furthermore, the generalized data mining tools are used for knowledge discovery in web, there are certain attempts at reviewing the website content mining and these were from the different perspective of the methods used and the issues solved but not in a sufficient complexity [4]. Mary, X. Leela, and Silambarasan. G presented an overview of data mining tools and techniques, and the authors highlighted the concept of web mining, and web content mining, the interrelationship of web mining, and various approaches such as structured web mining and web unstructured mining are also reviewed in their research work, the further analysis of the multiple tools and techniques is also provided in their research work [5]. Mughal, J.H Muhammad presented a review work on data mining, the concept of data is provided by the author with the preferences of web usage mining, web

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2024 doi:10.20944/preprints202407.2339.v1

structure and web content mining techniques to discover patterns of knowledge, and extract the relevant sources from huge amount of data from the WWW [6]. Thacker, Palak, and Thacker, Chintan presented a review work on web page ranking algorithms in the web mining, the collection of consistent websites and site pages provides essential information to the users. The gradual increase in the quantity of web pages functions as traditional PageRank algorithm requires numerous enhancement and adjustments in a versatile manner [7]. Mebrahtu, A, and Srinivasulu, B, presented an overview of web content mining techniques, tools and relevant tactics to provide spectacular and unpredictable progression of information present on the Internet. The authors have presented the notion of web mining, then provides a complete structure of techniques, strategies, and tactics of web mining and then provide a review of several types of web content mining tools, techniques, and strategic enhancements and complete with the relevant algorithms [8]. Vidya, S, and Banumathy, K, presented a review on web mining concepts, applications, and techniques, the authors explained the entire systematic approach of web mining, and it's subcategories that includes web content, web usage mining, and web structure mining along with the classification of clustering and association [9]. Siddiqui, A.T, and Aljahdali, S contributed in the research work on the production of E-Commerce Applications via Web Mining Techniques, the authors explained the notion of web mining techniques and their influence on the dimensions of businesses, enterprises, and companies [10]. Lang Chunmin and et al., provided an analysis for the examination of the consumer's fashion-oriented consumer exposures with the help of web content mining tactics and methodological approaches. The authors identified the merits and demerits of online mass customizations tailored to the experiences gathered by the individual customer or focus groups. Furthermore, cost analysis and estimation were predicted by the authors using different techniques of web content mining starting from the data cleaning process to data validation of clothes and vice versa [11]. Bhat Prashant and et al., proposed a new framework for social media analysis of content mining approaches and discovering patterns of knowledge. The authors reviewed some of the prevailing social media-based web content mining tactics to propose a new modular approach for effective data mining patterns and framework to extract the useful and manageable framework from the web usage data mining aspects [12]. Singh, Satyaveer, and Aswal, Singh Mahendra presented ontology-based learning approaches based on the classification of web mining techniques. The authors focused on the need for analysis of rapid and effective management of constructive ontologies for building a standardized knowledge-based and semantically driven web-based ontology software solution. Furthermore, the authors highlighted a comprehensive overview of various tasks in conjunction with the ontology-based learning methods and frameworks pertaining to the web mining tactics for comparatively analyzing distinct ontological learning-based tasks to overcome problems and provide solutions in devising the ontology from the semi-structured set of website pages [13]. Aartsen Van Brent, and et al., conducted a detailed review and analysis of web usage mining techniques and further prospective research initiatives. The authors reviewed the web usage techniques of specific years to identify and cultivate the initial state of web usage mining research capabilities to answer your research questionnaires and identify the sources of data used in the web usage mining and methods of data mining are capable of extracting the patterns of knowledge and data. The authors also classified the web usage mining applications and the futuristic research approaches that can be implemented in web usage mining based on the PRISMA method for conducting the personalization and recommendation-based systems using web usage mining techniques [14]. Jin Jingquan and Lin Xin proposed a security assessment model based on the aspects of data mining, the authors performed experimentation on safety and security measures to extract weblogs that can significantly affect the algorithms of data mining to extract the patterns of the web from the website servers, then you need to identify the major accessibility types or the user interests, and you need to a specific event as per the discovery of the patterns of users to identify the accessibility configuration and behavior of the user. The authors identified web log mining as a robust and streamlined data mining algorithm to identify the variety of logs embedded in the server deployed on the web and then understand the accessibility or interests and preferences and the user need to perform an identical state of the

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2024 doi:10.20944/preprints202407.2339.v1

preferences of user and patterns of your website's behavior for the verification of your security parameters and produce to concluding statement [15].

3. Approaches of Web Content Mining

Significantly, web content mining is a methodological technique that retrieves data from the web and effectively processes it to produce well-formed structure and arranges the data in a manner that searching for a required knowledge from the web services can be effectively done rapidly in a planned pace[4]. The overall dimensions of this approach constitutes of discovering structure data from web sources, identification, classification, and implementation of similar data is extracted. Furthermore, Figure 2 explains the classified approaches of web content mining in the section below [5].

3.1. Unstructured Data Mining

Content Mining can be performed on unstructured data including text mining of unstructured data provided with unidentified information. Similarly, the extraction of previously collected unknown information from various text sources is referred as text mining. Hence, the approachability of extraction of unstructured data in content mining requires techniques of data mining and text mining classifications which are as follows:

3.1.1. Information Extraction

The process of extracting information from the unstructured textual sources to allow finding entries along with classifying and storing them in a specific database. Significantly, the semantically improved information extraction, also regarded as semantic annotation merges such entities along with their semantic identities, descriptions and interconnections from a knowledge-based graph. With the incorporation of metadata to the concepts extracted from the discovery phase, this approach solves several challenges in the areas of enterprise content management and knowledge discovery.

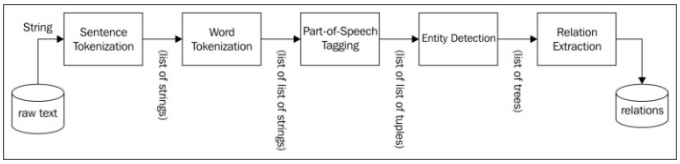

Figure 1.1. Block Diagram of Information Extraction.

3.1.2. Topic Tracking

The documents are related to the interests and preferences of the site users are predicated by examining the documents, the users' visits and by analyzing the user profiles. This technique is integrated by yahoo, users provides a keyword and if everything is interrelated to the keywords explodes then the users are acquainted about that specific objectives. Several areas of expertise including medicine, engineering, arts and education uses these methods to search contemporary progresses in their particular fields. The drawback of the strategic technique is that when users find for a specific topic then it'll be delivered to the admin with information which is not interrelated to the topic.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2024 doi:10.20944/preprints202407.2339.v1

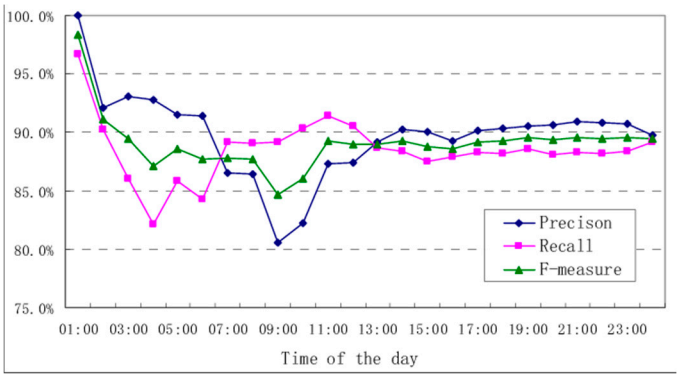

Figure 1.2. Graphical Representation of Topic Tracking.

3.1.3. Summarization

This techniques summarizes the complete document via maintaining the essential facts and points, helping users to adopt techniques to read and understand the topic or not. Similarly, the summarization practice uses two methodological approaches 1) extractive methodological approach, and 2) abstractive methodological approach.

- ✓ The extractive methodological approach chooses a subcategory of phrases, sentences, and words to formulate summary from the actual text.
- ✓ The abstractive methodological approach develops an internal semantic illustration and also uses NLP-based techniques to create summarization techniques. This summary would consists of words which aren't included in the actual textual document.

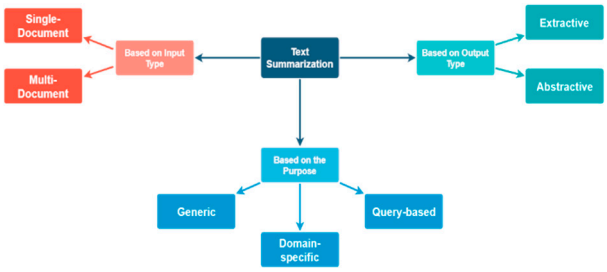

Figure 1.3. Illustration of Text Summarization and its classification.

3.1.4. Categorization

The sole aspect of this technique is to identify the major picture by inserting documents in a pre-built collection of groups. This strategic tactics computes the amount of words in the document and this evaluates, overall theme of document [8]. According to the considerations of rank topic is provided to the document. The overall documents comprises the mainstream contents on a specific topics at the initial ranks, this strategic technique helps to provide relevant client support to the business and multiple industries.

3.1.5. Clustering

The strategic technique is used to collect distinct documents, in this group of documents aren't accomplished on the concept of predefined topics, it is completed on the urgent basis. Although, some documents may seem in the dissimilar groups with a consequence of advantageous documents aren't lost from the search engine results. Hence, this method helps users to find topics of their interests and preferences.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2024 doi:10.20944/preprints202407.2339.v1

3.1.6. Information Visualization

Visualization uses feature extraction, key term indexing and relevant terms and conditions. The documents consisting of similarity of terms are found out via visualization. Significantly, the large quantity of textual materials are requested as the visualized maps or hierarchy where the facility of web browsing is allowed on the site pages. It facilitates users to analyze visual content, interact with proper scaling, zooming, and creating sub maps of the graphical representations.

3.2. Structured Data Mining

Structured content mining is a technique that is used to extract structured data from multiple websites or webpages [5]. The formulated in the list of data, tabular information and decisional tree are the instances of structured data forms. The key benefit of structured data is that it can be easily extracted as contrasted to unstructured data.

3.2.1. Web Crawler

A web crawler is a program or an automated scripts that browses the WWW in a methodological and automated manner. The process of visiting a website, reading site pages and relevant information to make records for indexing search engines is known as web crawling. Search Engines available onthe WWW comprises of similar programs, which is identified as a "bot", "spider", or "crawler. Search engines use crawlers often to collect available information on community-enabled websites or site pages. There are various types of crawlers that can be categorized in the form of internal and external web crawler. Hence, the working procedure of internal web crawler is that it crawls throughout the internal pages of the website and the external crawler traverses throughout the third-party or unknown websites.

3.2.2. Web Page Content Mining

The core focus of web page content mining is to classify web pages. It is a structured approach of web content mining, this methodological approach is operated by page ranking provided by the traditional search engines.

3.2.3. Wrapper Generation

The process of accumulating information from the wrapper generator on the competency of various sources is classified as wrapper generation. Significantly, the websites and web pages are hierarchical ranked by the search engines, by using the page rank factors the web pages can be easily retrieved based on the query.

3.3. Semi Structured Data Mining

Semi-structured data is the data which does not imitates to a data model but has certain structure, it lacks a fixed or rigid schema. It is a complex data that doesn't reside in a rational database but it has certain organizational properties that makes it a simple aspect to analyze it in a predictive manner. With certain aspect, users can store them in a relational database.

3.3.1. Object Exchange Model (OEM)

The appropriate information is usually extracted from semi-structured and is gathered in a collection of expedient information and is then stored in Object Exchange Model (OEM). This enables users to precisely examine the structural
information clusters that are accessible on the World Wide Web.

3.3.2. Web Data Extraction

This technique simply converts web data to structured data, the structured data is easily delivered to end users. Hence, the data is stored in the tabular form.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2024 doi:10.20944/preprints202407.2339.v1

3.3.3. Top Down Extraction

Top down extraction method solely emphasis on the extraction the composite objects from several web sources and converts them into a lesser amount of complicated objects until the actual objects are extracted.

4. Comparision of Web Content Mining Tools

The comparison of various web content mining tools requires a prior understanding and product awareness of the tools and technologies [7]. Web content mining tools are application software that assists users to download the important information, it usually collects suitable and detailed information. The following are some of the well-known data mining tools that are discussed below:

4.1. Web Info Extractor

Web Info Extractor is a versatile content mining and data extraction tool that is used for appropriate analysis of raw data, process and mines it in a well-structured form for contextual monitoring of content-based update [2][4].

4.2. Mozenda

Mozenda is a web scraping tool and service that is considered as 5 star rated customer support, it provides a cloud-hosted software, on premise software, and data services with 15 years of experience, and this allows users to regularly automate web content extraction from any site or platform.

4.3. Screen Scraper

Screen Scraper is an appropriate screen scrapping tool that searches a relational database, SQL server or database that can be connected with the backend software to successful achieve a web content mining configuration and relevant sources of information.

5. Results

The results obtained from the above discussed tools are highlighted in tabular form below:

Tool	Records Data	Extracts Structured Data	Extracts Unstructured Data	Easy to Use
Web Info Extractor	✓	✓	✓	✓
Mozenda	X	✓	✓	✓
Screen Scraper	X	✓	✓	X

6. Conclusions

Web Content Techniques are used to extract information from various sources across the Internet and the WWW. In this paper, exploratory mining tools and techniques are used to mine the information of web content on the internet, the overall analysis and theoretical overview is explained for the improvement of web content mining techniques to increase consistency, scalability, and rate of adaption to gradual increase the knowledge base and vice versa.

References

1. D. Eldhose T John, B. Skaria, and P. X. Shajan, "An Overview of Web Content Mining Tools," *Bonfring Int. J. Data Min.*, vol. 6, no. 1, pp. 01–03, 2016, doi: 10.9756/bijdm.8126.
2. N. R. Satish, "A Study on Applications , Approaches and Issues of Web Content Mining," *Int. J. Trend Res. Dev.*, vol. 4, no. 6, pp. 41–43, 2017.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2024 doi:10.20944/preprints202407.2339.v1

3. Z. Yu, "Optimization techniques in data mining with applications to biomedical and psychophysiological data sets," *ProQuest Diss. Theses*, vol. 1464852, p. 91, 2009.
4. M. O. Samuel, A. I. Tolulope, and O. O. Oyejoke, "A Systematic Review of Current Trends in Web Content Mining," *J. Phys. Conf. Ser.*, vol. 1299, no. 1, 2019, doi: 10.1088/1742-6596/1299/1/012040.
5. X. L. Mary and G. Silambarasan, "Web Content Mining : Tool , Technique & Concepts," vol. 7, no. 5, pp. 11656–11660, 2017.
6. M. J. Hamid Mughal, "Data mining: Web data mining techniques, tools and algorithms: An overview," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 6, pp. 208–215, 2018, doi: 10.14569/IJACSA.2018.090630.
7. P. Thacker, A. Prof, and C. Thacker, "a Review Paper on Various Web Page Ranking Algorithms in Web Mining," *Int. J. Adv. Eng. Res. Dev.*, vol. 3, no. 02, pp. 192–197, 2016, doi: 10.21090/ijaerd.030236.
8. A. Mebrahtu and B. Srinivasulu, "Web Content Mining Techniques and Tools," *Int. J. Comput. Sci. Mob. Comput.*, vol. 6, no. 4, pp. 49–55, 2017.
9. S. Vidya and K. Banumathy, "Web Mining- Concepts and Application," vol. 6, no. 4, pp. 3266–3268, 2015.
10. H. Gu, "Data mining in the application of e-commerce website," *Adv. Intell. Syst. Comput.*, vol. 180 AISC, no. 8, pp. 493–497, 2013, doi: 10.1007/978-3-642-31656-2_70.
11. C. Lang, S. Xia, and C. Liu, "Style and fit customization: a web content mining approach to evaluate online mass customization experiences," *Journal of Fashion Marketing and Management: An International Journal*, vol. ahead-of-print, no. ahead-of-print, Jul. 2020, doi: https://doi.org/10.1108/jfmm-12-2019-0288.
12. P. Bhat, P. Malaganve, and P. Hegde, "A New Framework for Social Media Content Mining and Knowledge Discovery," *International Journal of Computer Applications*, vol. 186, no. 36, pp. 17–20, Jan. 2019, doi: https://doi.org/10.5120/ijca2019918356.
13. S. Singh and M. S. Aswal, "Ontology Learning Procedures Based on Web Mining Techniques," *SSRN Electronic Journal*, 2019, doi: https://doi.org/10.2139/ssrn.3382660.
14. B. Van Aartsen, O. El-Gayar, and C. Noteboom, "A Systematic Review of Web Usage Mining Techniques and Future Research Options," *Research & Publications*, Jan. 2020, Accessed: Jul. 25, 2024. [Online]. Available: https://scholar.dsu.edu/bispapers/134/
15. J. Jin and X. Lin, "Web Log Analysis and Security Assessment Method Based on Data Mining," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–9, Aug. 2022, doi: https://doi.org/10.1155/2022/8485014.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.