International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 03, 2021

Web Mining And Web Content Mining -A Brief

Ms. A. Menaka[1][0000-0003-3804-0115], Dr. K. Sakthivel[2][0000-0002-7599-6258]
Mrs.K. Pugazharasi[3][0000-0001-6296-2174]

[1]*Assistant Professor of IT, Adhiyamaan College of Engg, Hosur*
[2]*Professorof CSE, K.S. Rangasamy College of Technology, Tiruchengode*
[3]*Assistant Professor of CSE, K.S. Rangasamy College of Technology, Tiruchengode*

Email: [1]asmenakaace@gmail.com,[2]sakthivelk@ksrct.ac.in,
[3]pugazharasi@gmail.com

Abstract: Data mining is the method of extracting useful knowledge from volume of data and converting it into usable framework for potential use. Data mining has various categories of applications as XML mining, web mining, text mining, etc. This paper gives a detailed concept of web mining. The growth of the web in recent years has increased gradually. There are billions of web pages, images, audio files, and video files available on the internet. Retrieving valuable user needed information from the web documents is a tedious process, and thus web mining has emerged. Web mining automatically extracts useful information from web documents based on a user keyword. Web mining has three categories used for mining the web: web content mining, web structure mining, and web usage mining.

Keywords: Data Mining, Web Content Mining, Web Structure Mining, Web Usage Mining, Web Mining, Internet.

1. INTRODUCTION

The World Wide Web (WWW) is an integrated system of internet-accessible public websites. Nowadays the WWW is getting more popular, and also it is a medium to access the variety of information stored on different websites in various web pages across the world. Data stored on the web is in formless. As the web's information expand daily at enormous rates, its tough to extract useful information based on user needs. Web mining is playing a significant role in resolving these issues. The problems when interacting with the web:

1.Finding the closely related information
2.Gathering knowledge from the data available on the www.
3.Customization of the knowledge
4.Intellect individual's users details

Web mining is a sub-disciplines of data mining related to the information accessible on the internet. It is a method to retrieve information accessible on different web pages on the internet. In today's scenario, different search engines are available.1Various tools, techniques, and algorithms are used to retrieve web pages that include various documents, images, advertisements, audio, video, etc. Web mining is swiftly becoming popular because web documents are rising on the internet every day. It is tough and long-lasting to find specific trends, expertise, and informational data if it is done manually [1]. Web data mining is cate-

gorized into three types: Web Content Mining(WCM), Web Structure Mining(WSM), and Web Usage Mining(WUM). WCM finds valuable information from web documents. WSM determines useful information that uses hyperlinks through which web pages link to each other.

Web usage mining examines and finds valuable information from user's log files. The log file contains user surfing details, registration details on the website, etc. All above mining uses various methods, algorithms to find knowledge on the www.

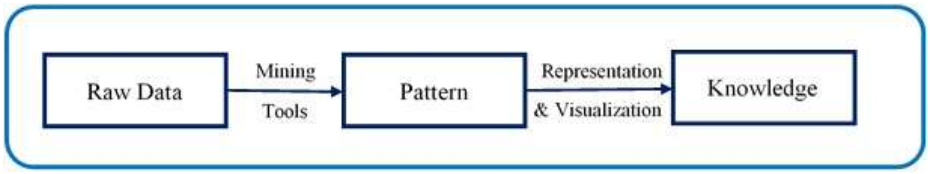

Fig. 1. Process of Web Mining

The following are sub-tasks to the web mining process [2]:
1.Finding the resource:It is the job of downloading planned web documents.
2.Selection and pre-processing of content: Dynamically sorting and pre-processing relevant site services from collected information.
3.Generalization:The particular website and various pages immediately discover general pattern.
4.Analysis: Validation and study of pattern extraction.

Categories

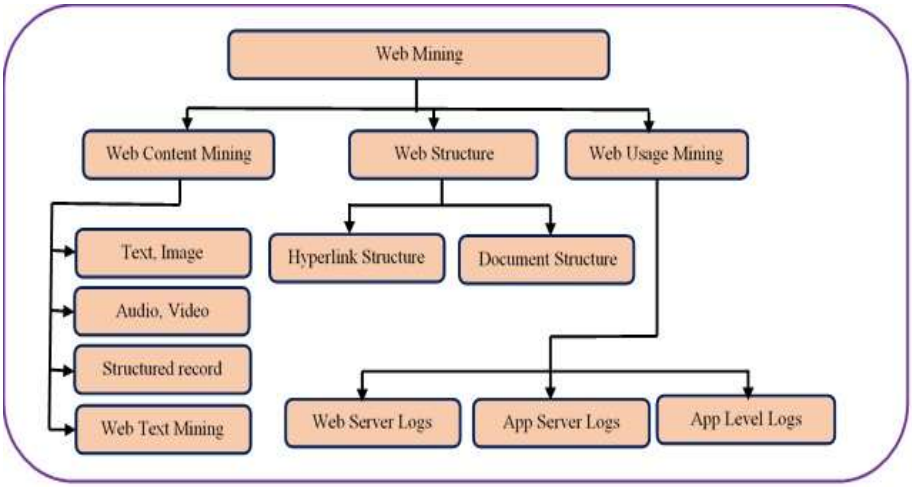

Fig. 2. Web Mining Categories

Web Information mining

Web information mining retrieves useful data from web records, such as text, videos, graphs and photographs. It is also known as text mining. In web information mining, there are two kinds of techniques used. The techniques are network approach and the agent-based [3]. Network method extracts the formless data from web pages. The agent-based method looks for specific data and helps to coordinate web sites using the gathered data. The material of online records is analyzed by web information mining. A web page includes a set of information that are composed to communicate information to the users.

Web Structure Mining

Method of collecting knowledge on the layout of web documents is referred to as web structure mining. It mines valuable knowledge from hyperlinks. The extracted knowledge illustrates the structure of the web page. Keyword search is a key application used to find relevant web pages. The web graph is composed of web pages and hyperlinks. The web pages are referred as nodes and hyperlinks are referred as edges that connect link between pages. The hyperlink that leads from other web documents into a single web page is called in-links. The hyperlinks created from the respective web documents are called out-links to other web pages.[4]. It is analyzed at two stages, the level of intra text and the level of inter hyperlink.

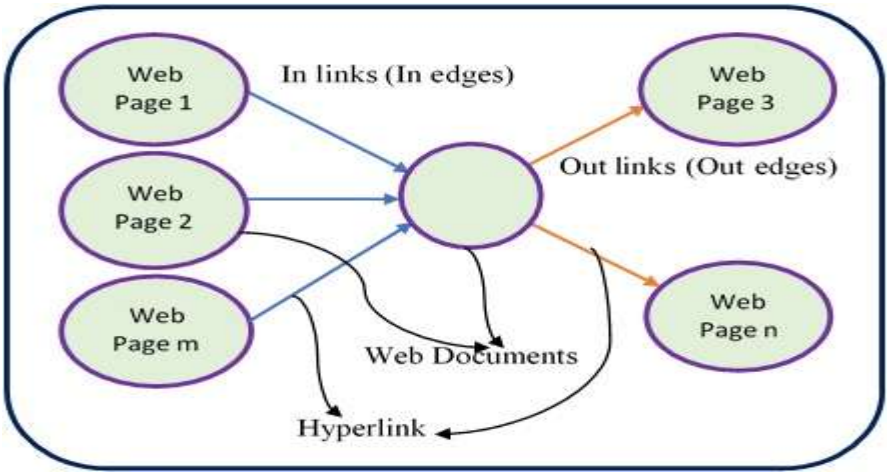

Fig. 3. Web Graph Structure

Web Usage Mining

It refers to mining in log data to retrieve user's performance in different applications like e-commerce, pre-fetching, modified services, etc. The log files provide site user operations, such as user browsing information. [5]. User behaviour logs are stored in the service log archives. Weblog data is naturally chaotic and mystify. The user usage data collected at web sites are IP address, the access time of web page, mouse clicks, no of scrolls, registration details, users queries, and other data the user used for interaction,

Brief Overview On WCM

Structured data, formless data, partially-structured and multimedia are grouped into four main groups. Any information stored on the site is unstructured. The view of the data retrieval view and the view of the database are two different perspectives of web mining content. The key goal of knowledge mining is to boost the collection and exploration of information from an information retrieval point of view. The primary objective of browsing the database is to monitor site data. The web content mining uses wrapper. It is a series of guidelines that are used to retrieve valuable facts from websites. Data stored on the web is text, audio, video, images, etc. The web information mining includes data sorting, data clustering, and attribute tagging for fast retrieval. This web information mining is linked to text mining; more than 50% of data are in text form. [6].

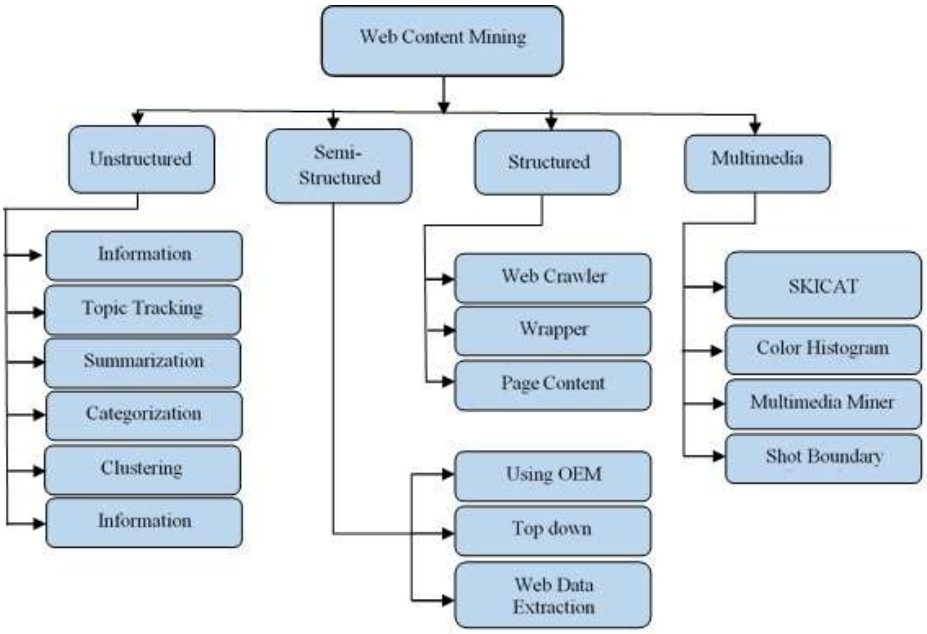

Fig. 4. Categorization of Web Content Mining

2. TECHNIQUES TO EXTRACT UNSTRUCTURED DATA

Information Extraction

It automatically derives the structured portion of the information from formless or partially-structured data.The initial point of process is to evaluate formless text.Pattern

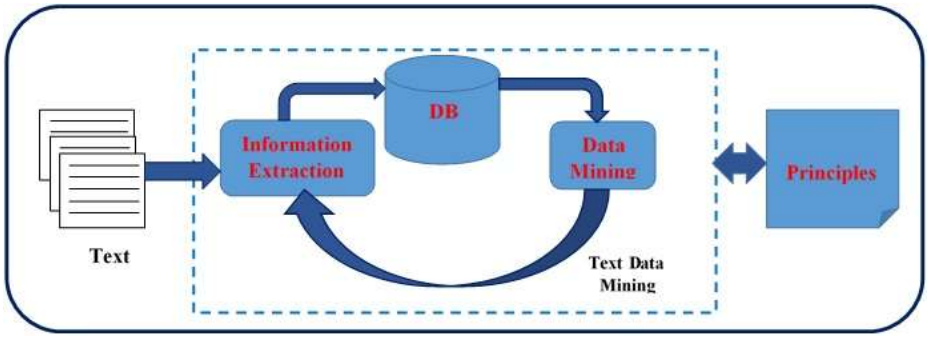

Fig. 5: Overview of IE-based text mining framework

matching is used to retrieve from formless data. Information extraction extracts collection of keywords in the text [6]. This method is useful to extract the important phrases from text. Information extraction provides the KDD module to transform unstructured data into structured data.

Topic Tracking

Topic monitoring is a method used to verify documents downloaded by users and to analyse their profiles. Based on the documents, user sees this methodology as defining other types of concern to the user. In the real-world, yahoo provides a free theme monitoring application (www.alerts.yahoo.com) helps users to select keywords and inform them when news becomes available. However, this technique has limitations. One sample example for topic tracking is if a user wants to alert himself/herself from competitor.

Similarly, business owners will want to monitor news about their own business, and goods may can use this technique. This topic tracking is used in the medical field by physicians and those searching for potential cures for disorders and having to keep up with the latest advances. Individuals in education may still use the subject monitoring to ensure that they have the current study references on their field of interest [6]. The limitation in topic tracking is this technique may give information that is not useful to the user interest.

Summarization

As the name Summarization defines that it is the process of reducing the length and meaning of the whole document in short and neat valuable points for the users. It also makes the user check if a long document suits their needs, and that reading document is worthy. In real life, one example of summarization is Microsoft Word's AutoSummarize function [7]. To further extracts the particular topic that the user is seeking is done by mixing summarization and customization. If the user needs a topic based on their relevance of information, the summarization tools can be applied for sorting out fast.

Categorization

It includes the task of defining the key concepts in the text. This method records the terms in the text and then finds the key subjects discussed in the documents. In categorization, documents with substantial material on a single subject are ranked first, and this method is used in different applications [7]. For example, the industry wants to provide assistance or response to consumer questions on a number of topics. If this method is used to identify a text based on topics, then end-users and customers will access better information they desire.

Clustering

The similar documents are assorted as one in clustering. This technique is different from the categorization in which the records are collected on the fly despite topics. In clustering, users can easily find the documents in which topics are there in scope and not in the user scope.

Information Visualization

Visualization inserts broad textual materials into a hierarchy or map and facilitates the browsing facilities [8]. Informatik V's Docminer tool shows the grouping of significant volumes of text and allows the user to examine data in document. User can communicate with text by zooming, scaling, etc. This technique is useful to identify criminal or detecting a case.

3. TECHNIQUES TO EXTRACT STRUCTURED DATA

Strategies used in structured mining are Web Crawler, Wrapper creation, and Page Mining.

Web Crawler

Crawlers or robots are heart of search engines. It continues to search on the network to locate every new web pages added to the web. The essence of the network is dynamically growing every day, it is a challenging task to navigate all those URLs on the web [8]. Various web crawlers are Focused Web Crawler(FWC), Distributed Web Crawler(DWC), Incremental Web Crawler(IWC), Hidden Web Crawler(HWC).FWC search pages based on a keyword-based approach and retrieve it which is relevant to keyword searches. It tries to recover the maximum count of keyword-relevant web pages on the web. The distributed web crawler uses numerous machines for indexing the content from the web documents. The incremental crawler is the method of revisiting standardised resource locators and prioritising them. The hidden web crawler, as the name describes the web sites, the hidden web crawler is extracted based on a keyword search for the web page and not accompanied by hyperlinks.

Wrapper Creation

The development of the wrapper provides data dependent based on the capability of the sources. The webpage is ranked on total number of times(count) the user visits the page [9]. The webpages are retrieved on user queries, and the resultant web page is pop-up based on page rank. The wrapper also provides various information like domains, statistics, etc. Wrappers can also be used to retrieve archives of search results from dynamically generated pages of search results retrieved by search engines.

Page Mining

It works on the search engines where the pages are ranked. It categories pages based on page rank. Based on the content of the page the page rank is given.

4. TECHNIQUES TO EXTRACT SEMI-STRUCTURED DATA

Semi-structured data does not conform to a data model but has some structure. Object exchange model(OEM), Top-down extraction(TE), and Web data extraction language(WDEL) are techniques of semi-structured data.

Object Exchange Model (OEM)

In OEM the data is retrieved from the semi-structured data and finally the extracted data is stored in object exchange model [9]. It is used to represent semi-structured data, and it allows people to understand more precisely the layout of knowledge on the site. It is ideally suited to a mixed and active environment.

Top-down Extraction

Top-down extraction technique removes abstract artefacts from data-rich web resources. Works by compiling an object content and the reference context definition to locate new objects [10]. Context details the essence of object. This approach works well in data, which presents some variations in their structure.

Web Data Extraction Language

The hypertext transfer protocol accesses the pages from World Wide Web(WWW)[14]. It fetches the page from the web and stores the pages into a central local database for future analysis through a web crawler. Thus web crawler plays a significant role in web data extraction[15]. Once stored in the local database, the contents of the page are parsed, searched, reformatted, and copied into a spreadsheet. It is used for contact scraping, price analysis, scraping of product review (to assess competition), weather monitoring, etc.

5. MULTIMEDIA DATA MINING TECHNIQUES

In a multimedia database, multimedia data mining allows to derive exciting information from multimedia datasets such as audio, video, images, graphics, voice, and a mixture of many types of datasets and stores in a multimedia database[16]. SKICAT, color Histogram Matching(HM), Multimedia Miner(MM), and Shot Boundary Detection(SBD) are techniques in multimedia data analysis.

Sky Image Cataloguing and Analysis Tool (SKICAT)

The Sky Image Cataloguing and Analysis Tool includes machine learning, expert system, and machine-assisted data to automatically measure sources in the sky to segregate as stars and galaxies[17]. This data used by an astronomer to perform research in scientific object analysis. It produces a digital catalog of sky objects [11]. Through the machine learning technique, the objects converted to usable human classes. Image processing is applied for the classification of a very large classification data set.

Multimedia Miner(MM)

Multimedia miner analyses various forms of knowledge using summarization, classification, and association in image databases, audio databases, and video databases.
Four major components of Multimedia Miner are
1. Image excavator-used to excavate the photographs and videos from the repository.
2. Pre-processor-used to excavate features of image and store data in the database.
3. Search kernel- use to match queries with features of photographs.
4.Discovery modules-Inorder to intelligently evaluate fundamental knowledge and patterns within images, image information mining routines are predominantly used.

Shot Boundary Detection

In content-based, the video shot boundary plays a significant role in video indexing and retrieval [12]. A single-camera takes photos at a given time. Recognizing the transitions from consecutive shots is referred to identification of boundary shots. Two major changes are sudden transformation and progressive change.

Comparative Analysis of Internet Information Extraction Methods:

The table below displays the WCM tools and activities the tools accomplish [13].

Table 1. Tools and their Respective Tasks

Tools	**Duties**			
	Data Archives	**Extract Unorganised Knowledge**	**Extract Organised Knowledge**	**Friendly to the Customer**
Mozenda	No	Yes	Yes	Yes
Screen Scraper	No	Yes	Yes	No
Web Info Extractor	No	Yes	Yes	Yes
Automation Anywhere	Yes	Yes	Yes	Yes
Web Content Extractor	No	Yes	Yes	Not for unorganised data

List of Difficulties Lookout in WCM are:

Information Extraction (IE)

Extraction of valuable data or information from the web content (web sites, web pages) and various documents (web resources) is a difficult extraction process on the web. Information extraction is the process of retrieving a massive volume of information from webpages on the internet.

Opinion extraction from online sources

Online users review from online shopping sites, blogs, forums, and chatrooms used for opinions, which is very important for marketing intelligence (advertisements, recommendations) and product benchmarking. Efficiently analyzing customer opinions on the web is a difficult task.

Knowledge synthesis

Manually processing the hierarchies is a time-consuming task. Only a few existing techniques are exploring the information redundancy on the web. The modern techniques are important to synthesize and arrange the pieces of web knowledge.

Segmenting Web pages and detecting noise

Nowadays, the web pages contain advertisements, navigation links to other sites/pages, copyright notices, etc. An exciting problem is removing the key material from the pages without irrelevant information.

6. CONCLUSION

In the present situation, a vast amount of data has been uploaded on the web every day. The web contains various forms of data like images, tables, text, videos, etc. It is essential to provide immediate action regarding the user query through faster retrieval of data from the web using an effective knowledge retrieval process. Here three consequential forms of web data mining strategies that aid to extract data are outlined. Web information mining is useful in extracting knowledge from documents containing texts, tables, images, etc. Web structure mining mechanism demonstrates the interaction between hyperlinks on the web pages that display the layout of web. Web usage mining is used to discover the thrilling use trends from user information.

7. REFERENCES

[1] Muhammad Jawad Hamid Mughal," Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview," International Journal of Advanced Computer Science and Applications(IJACSA), Vol. 9, No. 6, 2018.
[2] Darshna Navadiya and Roshni Patel, " Web Content Mining Techniques-A Comprehensive Survey," International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 10, December 2012.
[3] K. R. Srinath, "An Overview of Web Content Mining Techniques," International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 11 Nov - 2017.
[4] Claudia Elena Dinucă, "Web Structure Mining," Annals of the University of Petroşani, Eco
[5] nomics, 11(4), 2011, 73-84.2011.
[6] Bhupendra Kumar Malviya and Jitendra Agrawal, "A Study on Web Usage Mining: Theory and Applications," Fifth International Conference on Communication Systems and Network Technologies, 978-1-4799-1797-6,2015.
[7] Vishal Gupta and Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications," Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August 2009.
[8] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, " Tapping the Power of Text Mining," Communications of The Acm, September 2006, Vol. 49, No. 9. 2006
[9] Anish Gupta and Priya Anand, "Focused Web Crawlers and Its Approaches," International Conference on Futuristic trend in Computational Analysis and Knowledge Management, 978-1-4799-8433-6,2015.
[10] Faustina Johnson and Santosh Kumar Gupta, " Web Content Mining Techniques: A Survey," International Journal of Computer Applications (0975 – 888) Volume 47– No.11, June 2012.
[11] Berthier Ribeiro-Neto, Alberto H. F. Laender, and Altigran S. da Silva, "Top-down Extraction of Semi- Structured Data."1999.
[12] Deepika Bhadoria1, Dr. Pradeep Sharma, "Review Paper on Web Structure Mining," International Journal of Scientific Engineering and Research (IJSER), Volume 4, Issue 7, July 2016.
[13] D. S. Guru, Mahamad Suhil, and P. Lolika, "A Novel Approach for Shot Boundary Detection in Videos.", pp (209-220), 2014.

[14] Bharanipriya, V., and Prasad, K. "Web content Mining Tools: A Comparative Study. International Journal of Information Technology and Knowledge Management. Vol. 4. No 1,211- 215. 2011.
[15] Michael Azmy, "Web Content Mining Research: A Survey." pp (1-14),2005.
[16] Arvind Kumar Sharma, P.C. Gupta,"Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 1, Issue 8, October 2012.
[17] Krishnamoorthy, Sujatha, Changiiang Zhang, and Zhou Yanxin. "Implementation Of Image Fusion To Investigate Wall Crack." 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE). IEEE, 2020
[18] K.Venkatachalam, A.Devipriya, J.Maniraj, M.Sivaram, A.Ambikapathy, Iraj S Amiri, "A Novel Method of motor imagery classification using eeg signal", Journal Artificial Intelligence in Medicine Elsevier, Volume 103, March 2020, 101787