Review on Web Content Mining Techniques

Shipra Saini
Department of Computer Science & Engineering
Amity University, Noida, U.P

Hari Mohan Pandey
Department of Computer Science & Engineering
Amity University, Noida, U.P

ABSTRACT
Web data processing is the method of handling high volume of data. Previous research explains that handling/processing such data is not easy. Therefore, researchers utilize web mining, deals with identifying patterns, which user require. The second phase of web mining is known as web content mining, which dealt mining of pictures, text and graphs etc. The primary purpose of web content mining is to identify the relevance of content according to the queries. The focus of this paper is to present a detailed and comprehensive review of various methods applied for web mining/web content mining. The paper is divided into three parts to discuss, web content mining, web structure mining and web usage mining. Later, we presented application of these approaches for structured, unstructured, semi-structured and multimedia data mining techniques. The underlined motivation is to explore new possibilities in improving the existing techniques and identifying new ways/methods.

General Terms
Data Mining, Web Content Mining.

Keywords
Web content mining, structured data mining, unstructured data mining, semi-structured data mining.

1. INTRODUCTION
Internet is a shared global computing network. It enables global communication between all the connected computing devices. It is a platform for web services and World Wide Web [8]. Web is a popular and interactive medium with intense amount of data freely available for users to access. It is a collection of documents, text files, audios, videos and other multimedia data [2]. The different types of data have to be organized in such a way that different users can efficiently access it. Data mining means extraction of data in terms of patterns or rules from huge amount of data [1]. The term web mining was coined by Etzioni in 1996, to denote the use of data mining techniques to automatically discover web documents, extract information from web resources and uncover general patterns on the web. The research in the field of web is classified on two aspects: the retrieval and the mining. The retrieval focuses on retrieving relevant information from large repository whereas mining research focuses on extracting new information already existing data [3]. In past, techniques like information extraction, information retrieval and machine learning were used to discover new knowledge from huge amount of data available on web. Information extraction focuses on extracting relevant facts whereas information retrieval focus selects relevant document. Now, Web mining is a part of both information extraction and information retrieval. Web mining supports machine learning because it improves the classification of text [4]. The main aim of web mining is to extract information. Web mining is integration of information that is gathered by traditional data mining techniques with information gathered over World Wide Web. Web mining is decomposed into following subtasks [1]:

i. **Resource Discovery:** It helps in retrieving services and unfamiliar documents on web.
ii. **Information selection and preprocessing:** It automatically selects and preprocesses specific information from the web sources.
iii. **Generalization:** It uncovers general pattern at individual web sites as well as across multiple sites.
iv. **Analysis:** It validates and interprets the mined pattern.
v. **Visualization:** It presents the result in visual and easy to understand way.

Web mining is divided into three main categories depending on the type of data as web content mining, web structure mining and web usage mining [2].

2. WEB MINING CATEGORIES
This section divides web mining into three categories depending on the type of data i.e. Web Content Mining, Web Structure Mining and Web Usage Mining.

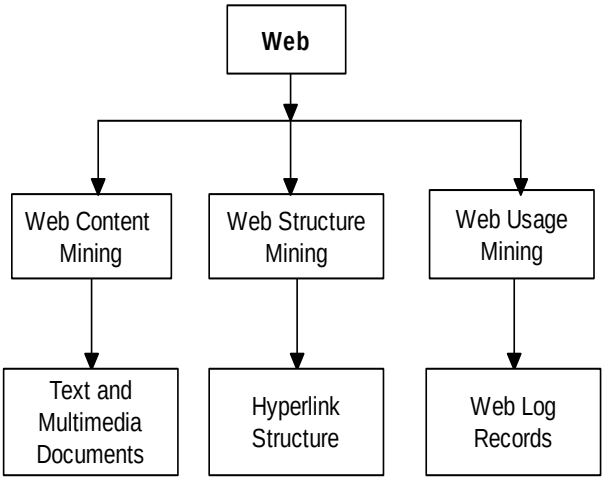

Figure 1: Web Mining Categories [5]

2.1 Web Content Mining
Web content mining is the mining, scanning and extraction of text, videos, graphs and pictures from web documents. It is also known as text mining. Two types of approaches are used in web content mining. The two approaches are: the database approach and the agent based approach. The database approach helps in retrieving the semi-structured data from web documents. The agent based approach searches relevant information and helps in organizing the collected information [6]. Web content mining analyzes the content of web resources. Content data correspond to collection of facts a web page was designed to convey to the users. Most of the data available on the web is unstructured data. Two different points of view of web content mining are: the information retrieval view and the database view. The main goal of

content mining from information retrieval view is to improve the filtering and finding of the information to the users. The main goal of database view is to manage the web data [7].

2.2 Web Structure Mining

The goal of web structure mining is to generate structural summary about web pages and web sites. It shows the relationship between the user and the web. It discovers the link structure of hyperlinks at the inter document level [6, 8]. It also helps in discovering the structure of document which is used in revealing the structure the structure of web pages and it's possible to compare the web page schemes [9]. This is further divided into two types that is based on the kind of structural information used [10].

a) **Hyperlinks:** Hyperlinks help in connecting web pages to different location either in same web page or on different web page. A hyperlink is divided into two categories i.e. intra-document hyperlink and inter-document hyperlink. Intra-document hyperlink connects different part of the same page whereas inter-document hyperlink connects two different pages.
b) **Document Structure:** The content within the web page can be organized in tree structure that is based on various HTML and XML tags.

2.3 Web Usage Mining

Web usage mining is the process of finding out what users are looking for on the internet. It tries to discover useful information secondary data derived from the interaction of users while surfing web [8]. There are three phases of web usage mining. The three phases are [9]:

a) **Preprocessing:** It helps in retrieving the raw data from web resources and then processes the data.
b) **Pattern Discovery:** After preprocessing the data, the data is used for discovering patterns.
c) **Pattern Analysis:** After discovering the pattern the pattern is analyzed and then the pattern is checked. If the pattern is correct then it is implemented on web to extract the information from web.

3. APPROACHES OF WEB CONTENT MINING

Web Content mining has following approaches to mine data: unstructured mining, structured mining, semi-structured mining and multimedia mining.

3.1 Unstructured Data Mining

Text document is the form of unstructured data. Most of the data that is available on web is unstructured data. The research of applying data mining techniques to unstructured data is known as knowledge discovery in texts [4].

3.1.1 Information Extraction

To extract information from unstructured data that is present on web pattern matching is used. It traces the keywords and phrases and then finds out the connection of keywords within text. When large volume of text is there then the technique is very useful. Information extraction transforms unstructured text to more structured form. First, from extracted data the information is mined, then using different types of rules, the missed out information is found. Information extraction making incorrect predictions on data is discarded [5, 7].

3.1.2 Topic Tracking

This technique checks the documents viewed by the user and studies the user profile. It predicts the documents related to users interest. The topic tracking applied by yahoo, user give a keyword and if anything related to keyword pops then the user is informed about that. This technique can be applied by many fields. The two fields where it is used is medical field and education field. In medical field doctors easily come to know about the latest treatments. In education field it is used to find out the latest reference for research related work. The disadvantage of the technique is that when we search for our topic then it may provide us with information which is not related to our topic [3, 5].

3.1.3 Summarization

The technique is used to reduce the length of the document by maintaining the important points. It helps the user to decide whether to read the topic or not. The time taken by the technique to summarize the document is less than the time taken by the user to read the first paragraph [5]. The summarization technique uses two methods that is the extractive method and the abstractive method. The extractive method selects a subset of phrases, sentences and words to form the summary from the original text. The abstractive method builds an internal semantic representation and then uses natural language generation technique to create the summary. This summary may contain words which are not present in the original document [3].

3.1.4 Categorization

This technique identifies the main theme by placing the documents in a predefined set of group. The technique counts the number of words in the document and this decides the main topic. According to the topic the rank is given to the document. The documents with majority contents on particular topic are given first rank. This technique helps in providing customer support to the industries and business [5] [7].

3.1.5 Clustering

The technique is used to group similar documents. In this grouping of documents is not done on the basis of predefined topics. It is done on fly basis. Some documents may appear in different group. As a result useful documents are not omitted from search results. This technique helps user to select the topic of interest [5].

3.1.6 Information Visualization

Visualization utilizes feature extraction and key term indexing. Documents having similarity are found out through visualization. Large textual materials are represented as visual maps or hierarchy where browsing facility is allowed. It helps in visually analyzing the content. The user can interact by scaling, zooming and creating sub maps of the graphs [4].

3.2 Structured Data Mining

The techniques are used to extract structured data from web pages [11]. Data in the form of list, tables and tree is structured data. The structured data is easy to extract as compared to unstructured data.

3.2.1 Web Crawler

Crawlers are computer programs which traverse the hypertext structure in web. Web crawlers can be used by anyone to collect information from the web. Search engines use crawlers

frequently to collect information about what is available on public web pages [3]. There are two types of crawlers. They are internal and external web crawler. Internal web crawler crawls through internal pages of the website and the external crawler crawls through unknown websites [5].

3.2.2 Page Content Mining

Page content mining is a technique that is used to extract structured data which works on the pages that are ranked by the traditional search engines. The pages are classified by comparing the page content rank [4].

3.2.3 Wrapper Generation

The information is provided by the wrapper generator on the capability of sources. Web pages are ranked by traditional search engines. By using the page rank value the web pages are retrieved according to the query [4].

3.3 Semi-Structured Data Mining

Semi-structured data arises when source does not impose rigid structure on data. If we want to extract data from web page and populate that data in database [3].

3.3.1 Object Exchange Model

The relevant information is extracted from semi-structured and is collected in a group of useful information and is then stored in Object Exchange Model (OEM). This helps the user to accurately understand the structure of the information that is available on web. The main feature of the model isthat it is self-describing that is there is no need to describe the structure of an object in advance [4, 7].

3.3.2 Top down Extraction

This technique helps in extracting complex objects from a rich web sources and decompose them into less complex objects until atomic objects have been extracted [5].

3.3.3 Web Data Extraction Language

This technique helps in converting web data to structured data and then delivers this data to end users. The data is stored in the form of tables [4].

3.4 Multimedia Data Mining

Multimedia data mining is the process of finding interesting patterns from media data such as video, audio, text and images that are not accessible using queries [3].

3.4.1 SKICAT

SKICAT is a successful astronomical data analysis and cataloging system that produces digital catalog of sky object. It uses machine learning techniques to convert the objects to human usable classes. It integrates technique for image processing and data classification which helps to classify very large classification set [7].

3.4.2 Color Histogram Matching

Color histogram matching consists of color histogram equalization and smoothing. Equalization tries to find the correlation between color components. The problem faced by equalization is the presence of unwanted artifacts in equalized images. The problem is solved by using smoothening [4].

3.4.3 Multimedia Miner

Multimedia miner consists of four major steps. Image excavator for extraction of images and videos, a preprocessor for extraction of image features and are stored in a database. A search kernel is used for matching queries with image and video that are available in database. The discovery modules mine image information to trace out the patterns in image [5].

4. CONCLUSION

Data mining techniques are used for the extraction of web information. Large amount of data is maintained by the web sources and that data can be clearly extracted by web mining techniques when the techniques are used accurately according to the requirements of the user [12]. Web content mining has been proved very useful in the business world. Users find difficulty in deciding which information is relevant to them from general purpose search engines. Web content mining solves the problem and helps the users in fulfilling their needs [4]. The main purpose of web content mining is to gather, organize, categorize and provide the user with the best possible information that is available on World Wide Web [5]. The detailed study and analysis of each web content mining technique have been done in this paper. The future scope of web content mining is to predict the user needs to improve the usability and scalability.

5. REFERENCES:

[1] Singh, Brijendra, and Hemant Kumar Singh.“Webdata mining research: A survey." Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on. IEEE, 2010.

[2] R. Malarvizhi and K. Saraswathi.“Web Content Mining Techniques Tools & Algorithms-A Comprehensive Study." International Journal of Computer Trends and Technology (IJCTT), Volume 4, 2013.

[3] Deepti Sharda and Sonal Chawla. "Web Content Mining Techniques: A Study." International Journal of Innovative Research in Technology & Science.

[4] Johnson, Faustina, and Santosh Kumar Gupta. "Web Content Mining Techniques: A Survey."International Journal of Computer Applications (0975–888) Volume (2012).

[5] Sharma, Arvind Kumar, and P. C. Gupta. "Study & Analysis of Web Content Mining Tools to Improve Techniques of WebData Mining." International Journal of Advanced Research in ComputerEngineering & Technology (IJARCET) Volume 1 (2012).

[6] Hussein, Mohamed-K, and Mohamed-H, Mousa. "An Effective Web Mining Algorithm using Link Analysis." (IJCSIT) International Journal of Computer Science and Information Technologies 1.3 (2010).

[7] Srividya, M., D. Anandhi and M. I. Ahmed. "Web mining and its categories–a survey."International Journal of Engineering and Computer Science, IJECS 2.4 (2013).

[8] Manoj Pandia, Subhendu Kumar Pani and Sanjay Kumar Padhi. "A Review of Trends in Research on Web Mining." International Journal of Instrumentation, Control and Automation, Volume 1, 2011.

[9] Sharma, Kavita, Gulshan Shrivastava and Vikas Kumar. "Web mining: Today and tomorrow." Electronics Computer Technology (ICECT), 2011 3rd International Conference Volume 1, IEEE, 2011.

[10] Srivastava, Prasanna Desikan and Vipin Kumar. "Web mining–concepts, applications and research directions." Foundations and Advances in Data Mining, Springer Berlin Heidelberg, 2005.

[11] Pol and Kshitija. "A Survey on Web Content Mining and extraction of Structured and Semi-structured data." Emerging Trends in Engineering and Technology, 2008, ICETET'08, First International Conference ,IEEE, 2008.

[12] Ananthi J. "A Survey Web Content Mining Methods and Applications for Information Extraction from Online Shopping Sites." International Journal of Computer Science & Information Technologies 5.3 (2014).

[13] Dohare, Mahendra Pratap Singh, Premnarayan Arya, and Aruna Bajpai. "Novel Web Usage Mining for Web Mining Techniques." International Journal of Emerging Technology and Advanced Engineering 2.1 (2012).

[14] Liu, Bing, and Kevin Chen-Chuan-Chang. "Editorial: special issue on web content mining." Acm Sigkdd explorations newsletter 6.2 (2004).

[15] Kosala, Raymond, and Hendrik Blockeel. "Web mining research: A survey." ACM Sigkdd Explorations Newsletter 2.1 (2000).