ISSN:2321-1156

International Journal of Innovative Research in Technology & Science(IJIRTS)

# WEB CONTENT MINING TECHNIQUES : A STUDY

Deepti Sharda, Assistant Professor, Department of Computer Science & Applications, MCM DAV College for Women, Chandigarh;
Sonal Chawla , Chairperson, Department of Computer Science, Punjab University, Chandigarh

## Abstract

Web Mining is extracting information from the web resources and finding interesting patterns that can be useful from ever expanding database of World Wide Web. One of the subfield of Web mining is Web Content Mining. The objective of this paper is four folds. 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.

Keywords: Web Content Mining, Unstructured data, structured data, Semi-Structured data, Multimedia Data Mining Literature Review

## Introduction

With billions and billions of web pages available on World Wide Web, it is eventually turning into rich knowledge database. The knowledge does not come only from the contents of the web pages but also from the unique feature of Web, its hyperlink structure and the diversity of contents. Analysis of these characteristics often reveals interesting patterns and new knowledge which can be helpful in increasing the efficiency of the users. Extraction of knowledge is a challenging problem because of many factors such as size of the Web, its unstructured and dynamic content as well as its multilingual nature. Furthermore, the Web generates a large amount of data in other formats that contain valuable information e.g. Web server logs information about user access patterns can be used for information personalization or improving Web page design. The various activities and efforts made in this direction are referred to as Web mining.

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. Web mining research overlaps with other areas like artificial intelligence along with machine learning techniques, data mining, informational retrieval, text mining and Web retrieval. Over the years, advancements have been made in this field and many data mining techniques have been formed to discover resources, pattern and knowledge from the Web and Web related data (such as Web usage data or Web server logs).

## Overlapping Areas:

Web mining research overlaps substantially with other areas but mainly the research made in the field of Web can be classified on the basis of two aspects: the retrieval and mining. Retrieval research focuses on retrieving relevant information resources from large repository, while mining research focuses on discovering new information from existing data. First of all relevant documents are retrieved using Information retrieval, then relevant facts are extracted out of these relevant documents using information extraction. The next step is the use of machine learning techniques and data mining techniques to generalize this data and in the last step analysis is being made of these new mined patterns.[1]

Web mining deals with three main areas: web content mining, web usage mining and web structure mining.

Web content mining focuses on techniques for assisting a user in finding documents that meet a certain criterion (text mining). It extracts or mines useful information or knowledge from web page contents. It is similar to traditional data mining but it can also discover patterns in Web pages to extract useful data such as descriptions of products, postings of forums, etc, for many purposes. Furthermore, we can mine customer reviews and forum postings to discover consumer sentiments. These are not traditional data mining tasks.

Web structure mining aims at developing techniques to identify quality of web page which can be find out with the help of hyperlinks. For example, from the links, we can discover important Web pages, which, incidentally, is a key technology used in search engines. We can also discover communities of users who share common interests. Traditional data mining does not perform such tasks because there is usually no link structure in a relational table.

Web usage mining focuses on techniques to study the user behavior when navigating the web. It is also known as Web log mining. It refers to the discovery of user access patterns from Web usage logs, which record every click made by each user.

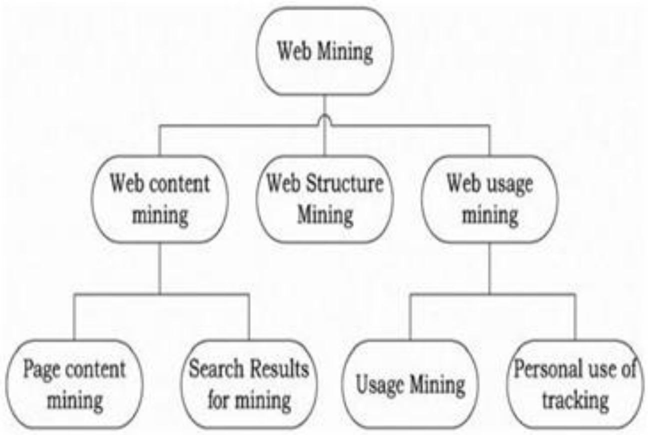

**Fig 1 Web Mining Taxonomy**

**Table 1 Web Mining Areas**

| Type of Web Mining | Web Content Mining | Web Structured Mining | Web Usage Mining |
|---|---|---|---|
| Type of data | Text Documents, Hyperlink Structures | Link structures | Server logs, Browser Logs |
| Representation | Group of words, Terms, Phrases, Concepts, Relational | Graph | Relational Table, Graph |

This paper mainly discusses Web content mining in detail as follows:

## 2.0 Web Content Mining

Web Content mining refers to the discovery of useful information from the contents of the webpage using text mining techniques. Webpage can be in traditional text form or in the form of multimedia document containing table, form, image, video and audio. Web content mining identifies the useful information from the Web contents. However, such a data in its broader form has to be further narrowed down to useful information.

Web Content mining relies a great deal on data mining and text mining techniques but all of its techniques are not based on them. Actually it is the subfield of data mining but it is not subfield of text mining. Text mining refers to the process of analyzing unstructured textual data, extract numeric indices from the text and thus make the information accessible to various data mining algorithms. In case of Web mining not all data is in textual form, it also involves non textual data such as Web server logs and transaction based data. The Web Content data can be in 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. Different techniques need to be applied in all three cases.[2]

## Unstructured Text Data Mining:

Web content data is much of unstructured text data. Data Mining techniques that are applied to unstructured data is termed as Knowledge Discovery in Text (KDT), or text data mining or simply text mining. Text Mining is a subset of the domain of data mining techniques. Retrieval of information from HTML web pages in itself is a challenging task. This is due to the fact that HTML web pages have multiple tags which are required to identify information and secondly because the web pages are highly unstructured. The rich variety of tags may pose a problem in case they are not processed correctly. We have to use some tools or techniques to get relevant data or information from that data. Next we will discuss techniques used for unstructured data mining.

## Topic Tracking:

It is the technique which is used to track the interest of the user. It checks the documents viewed by the user and tries to locate other related documents. Topic tracking is generally used by registered sites. For e.g. Yahoo uses this technique. The advertisements that are displayed when you login are related to the subject of mails you are receiving. Topic tracking can be beneficial in the field of medicine and research also. Any advancement done anywhere in the world can be notified to the registered user. Individuals in the field of education could also use topic tracking to be sure they have the latest references for research in their area of interest. It can be beneficial in the field of business also. A company can keep a check on its competitor by analyzing all the news that appear about their competitor. Topic tracking helps to track all subsequent stories in the news stream. Disadvantage of topic tracking is that sometimes we are not provided by the desired information. We may be provided with off track information [2].

## Summarization:

Summarization is the technique used to reduce the length of the document or that of multiple documents into a short set of words or paragraph that conveys the meaning of the text. It helps the user to decide whether the topic is of his interest or not. Two methods are used for summarization they 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 that is closer to what a human might generate. This kind of summary might contain words not explicitly present in the original document. Extractive methods for summarization are mostly used as compared to that of abstractive methods. Extractive methods search for headings and sub headings to find out the important points of that document. Summarization technique can work along with Topic tracking.[3]

## Categorization:

Categorization is the technique of categorizing the document. First of all it counts the number of words in a document then it decides the main topic from the count. It ranks the document according to the topic. Documents having majority content on a particular topic are ranked first. [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. There are various clustering algorithms available which can help user to easily select the topic of interest from the best relevant grouping.

## Structured Text Data Mining:

Structured data are typically the data records retrieved from underlying database and displayed in the web pages. It can be displayed either as tables or forms. Data can be extracted from these sources using structured data extraction techniques. This can be helpful in making value aided services by collecting information from various sources e.g. customized Web information gathering, comparative shopping, meta-search.
Following techniques are used for mining structured data:

## Web Crawler:

A web crawler is a relatively simple automated program, or script that methodically scans or "crawls" through Internet pages to create an index of the data it's looking for; these programs are usually made to be used only once, but they can be programmed for long-term usage as well. There are several uses for the program, perhaps the most popular being search engines using it to provide web surfers with relevant websites. Other users include linguists and market researchers, or anyone trying to search information from the Internet in an organized manner. Alternative names for a web crawler include web spider, web robot, bot, crawler, and automatic indexer. Crawler programs can be purchased on the Internet, or from many companies that sell computer software, and the programs can be downloaded to most computers. There are various uses for web crawlers, but essentially a web crawler may be used by anyone seeking to collect information out on the Internet. Search engines frequently use web crawlers to collect information about what is available on public web pages. Their primary purpose is to collect data so that when Internet surfers enter a search term on their site, they can quickly provide the surfer with relevant web sites. Linguists may use a web crawler to perform a textual analysis; that is, they may comb the Internet to determine what words are commonly used today. Market researchers may use a web crawler to determine and assess trends in a given market.

## Wrapper Generators:

To facilitate effective search on the World Wide Web several Meta Search Engines have been formed which do not do the search themselves but take help of the available search engines to find the required information. Meta Search Engines are connected to search engines by the means of Wrappers. For every search engine connected to it, there is a wrapper which translates user’s query into native query language and format of the search engine. The wrapper also extracts the relevant information from the HTML result page of the search engine. [5]

## 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 e.g. Bibliographic data where some books are written by single author and some by two or more authors. If we have to extract data from the web page and populate it in database then semi structured data mining techniques are required. Web pages provide some inherent structure which can be readily recognized but still one web page can differ from other web page significantly so we say the data of web page is semi structured.[6] In case of structured data we can extract data by submitting queries but it becomes difficult to query text data. To extract the data, we need some description of what to extract. Following techniques can be applied for extracting data from semi structured data:

## Top Down Strategy:

Using Top down strategy we can extract complex objects by decomposing them into less complex objects until atomic objects have been extracted. Through this technique just a couple of examples are sufficient for extracting hundreds of objects on a new web page. The central idea of this approach is to find out objects identical to the object we are considering. This object that we are considering is supplied by the user and it is very crucial as whole extraction procedure depends on this example object. Top down strategy works by traversing the structure of example object in preorder form visiting all its components and concatenating them to form new resultant object. Each new object is recognized and extracted in its entirety prior to identification of its component objects. [7]

## Wrapper:

This technique uses OEM (Object Exchange Model) which is particularly well suited for representing semi structured data. The wrapper uses the extractor to retrieve the relevant data in OEM format, and then executes the query (or whatever query conditions have not been applied) at the wrapper. The client receives an OEM answer object, unaware that the data was not stored in a database system. [8]

## NLP Techniques:

Data can be extracted from web sources using NLP (Natural Language Processing). NLP techniques are used to find out relevant fragments that can be extracted from source document. [9]

## TINTIN(Table Information-based Text Inquiry):

This tool extracts tabular data from unstructured documents based on a purely structured analysis of such documents. [10]

## Multimedia Data Mining Techniques:

Multimedia data mining (MDM) can be defined as the process of finding interesting patterns from media data such as audio, video, image and text that are not accessible using queries. MDM is the mining of knowledge and high level multimedia database system. Text mining, Image mining, Audio mining, and Video mining come under Multimedia data mining.

## Image Mining:

Image processing has been around for quite a while. Image processing focuses on detecting abnormal patterns as well as retrieving images. Image mining is all about finding unusual patterns. It deals with making association between different images present in large database.

## Video Mining:

Mining video data is more complicated than image mining because here we have collection of moving images in the form of animation. Video mining involves finding association between video clips and to find out unusual pattern in video clips.

## Audio Mining:

Audio data consists of 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. It can also be mined directly by using audio information processing techniques and then mining selected audio clips. [11]

## Analysis:

Heterogeneity of data on web poses difficulty to have one generalized technique to extract information. Therefore various Web content mining techniques are required to extract information from the web, based on the type of data that we are dealing with. As discussed above data can be in unstructured, structured, semi-structured and multimedia form. Each form requires different techniques. Techniques used for unstructured data are analyzed in the following table:

**Table 2 Unstructured Data Techniques**

| Technique | Topic Tracking | Summarization | Categorization | Clustering |
|---|---|---|---|---|
| Input | Database | Document | Document | Document Collection |
| Method | Topic Filtering from database<br>Topic Re-Ranking<br>Topic tracking | Extractive Method<br>Abstractive Method | Count the words<br>Explore main theme<br>Ranking | Categorization<br>Exploring best group |
| Output | Related Topics | Short set of words that conveys the message of the document | Related document according to ranking | Best relevant group containing desired document |

Techniques required for structured data available on web are analyzed in the following table:

**Table 3 Structured Data Techniques**

| Technique | Web Crawlers | Wrappers |
|---|---|---|
| Method | Crawls through hyperlinks.<br>Creates Index Database | Connect various search engines.<br>Passes the query from one search engine to another.<br>Resolves formatting issues existing between Search Engines |
| Use | Search Engines<br>Textual Analysis<br>Assess Market Trends | Meta Search Engines |
| Example | Google<br>Altavista<br>Yahoo | Visual Web Ripper<br>iMacros<br>Screen-Scraper Enterprise |

Techniques required for semi-structured data available on web are analyzed in the following table:

**Table 4 Semi-Structured Data Techniques**

| Technique | Top Down Strategy | Wrapper | NLP Technique | TINTIN Technique |
|---|---|---|---|---|
| Method | Extract Complex Objects<br>Decompose complex objects<br>Extract atomic objects | Connect various search engines.<br>Passes the query from one search engine to another.<br>Resolves formatting issues existing between Search Engines | Requires previous knowledge of data source.<br>Additional work required when source changes. | Extracts tabular data from unstructured documents.<br>Perform structured analysis of these documents. |
| Advantage/ Disadvantage | Simple<br>Effective | Requires previous knowledge of data source.<br>Additional work required when source changes. | Effective<br>Requires training example | Heuristics are required. |

## Conclusion:

From the above analysis of various techniques the paper concludes that different techniques are required as data available on web is not homogeneous. Under unstructured data, topic tracking technique is useful in tracking the interest of the user by locating other related documents. Summarization technique helps the user to decide whether they should read a particular topic or not. Categorization technique performs ranking of related documents whereas Clustering is helpful in finding the best relevant group in which desired document exists. These techniques can be used in combination also like Topic Tracking can be used along with summarization.

Web crawlers, technique used for structured data, is quite helpful for various search engines as they create index database. Wrappers are useful in case of Meta Search Engines.

Out of all techniques used for semi-structured data, Top Down Strategy is simple and effective. Wrappers can also be used for semi-structured data. Wrappers require previous knowledge of various search engines and additional work is required for connecting various search engines. NLP Technique is effective but requires training examples to work properly. TINTIN Technique is helpful if structured analysis is to be done of unstructured data.

Apart from being unstructured, structured or semi structured the data present on web can be in any form. It may be in text form, audio form, image form or in video form. This gives rise to Multimedia Data Mining Techniques. All these mining techniques are in infancy stage. These areas need to be explored so as to cater ever increasing and demanding needs of the users.

# References

[1] Kosla, R. and Blockeel, H. 2000. "Web Mining Research: A Survey." SIG KDD Explorations. Vol. 2, 1-15.

[2] 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.

[3] Rafeeq Al-Hashemi,Al-Hussein Bin Talal "Text Summarization Extraction System (TSES) Using Extracted Keywords."

[4] Roxana K. Aparicio Carrasco, "Semi Supervised Classification of Web Content using Mixture Models."

[5] Boris Chidlovskii, Jon Ragetli, and Martin de Rijke "Automatic Wrapper Generation For Web Search Engines."

[6] S .Abieboul. "Querying Semi-structured Data" In Proceedings of International Conference on Database Theories, pages1-18, Delphi, Greece, 1997.

[7] Berthier Ribeiro-Neto Alberto H.F.Laender Altigran Da Silva "Top Down Extraction of Semi-Structured Data."

[8] J. Hammer, H. Garcia-Molina, J. Cho, R. Aranha, and A. Crespo "Extracting Semistructured Information from the Web."

[9] J.Cowie and W. Lehnert. "Information Extraction", Communications of the ACM,39(1):80-91,1996.

[10] P. Pyreddy and W.B.Croft. "TINTIN: A System for Retrieval in Text Tables." In Proceedings of the Second ACM International Conference on Digital Libraries, pages 193-200,1997.

[11] Bhawani Thuraisingham, Managing and mining Multimedia Databases, International Journal on Artificial Intelligence Tools,Vol. 13,No. 3(2004), 739-759