International Journal of Computer Applications Technology and Research
Volume 4– Issue 3, 209 - 213, 2015, ISSN:- 2319–8656

Web Content Mining equipped Natural Language Processing for handling web data

Karan Sukhija
Research Scholar
Department of Computer Science and Application
Panjab University, Chandigarh
rs.karansukhija@gmail.com

Abstract: The growing usage of the web has unfolded the Web mining technology to a great extent. Web mining helps in extraction of useful knowledge from web data. (i.e. a range of web pages, hyperlinks among various pages, web sites usage logs and so on. This paper has threefold aspect. Firstly, it defines how web mining research area focuses on mining research and retrieval research (i.e. retrieval of data, information on web, data and text mining). Secondly, it categorizes the Web mining as content mining (i.e. retrieval of information from texts, images and other contents), structure mining (i.e. finding of facts from association of web pages) and usage mining (i.e. mining of information about usage of web sites). Web content mining mainly focuses on the structure of inner-document whereas web structure mining aim is to discover the linkage assembly of the hyperlinks at the inter-document level. Web usage mining includes three major phases i.e. preprocessing, pattern discovery and pattern analysis. Thirdly, it focuses on natural language processing as a backbone for web content mining that helps in handling of unstructured data over the web by offering various techniques. This paper concluded the web mining as trending research area for various research communities such as Databases, Artificial intelligence, Information retrieval and E-commerce.

Keywords: Web mining, Content mining, Structure mining, Usage mining, Opinion mining, Natural language processing.

1. Introduction

Web mining is promising trend of data mining that helps in finding of useful facts from web data. Web data includes the various web documents, hyperlinks among pages, web sites usage log and so on. Web mining can be defined either by the process-centric view approach or by the data-centric view approach. In process-centric view, web mining is defined as a sequence of tasks whereas in data-centric view, it is defined by means of web data that helps in the mining process [1]. Web mining research area focuses on mining research (i.e. discover hidden facts) and retrieval research (i.e. retrieves existing data or documents from a large database or document repository). Table 1 summarizes the possible categorization of retrieval and mining. This categorization is founded on twofold phases: Purpose and sources of data. The purpose of data retrieval techniques is to enhance the fetching of data from a databank and data mining techniques is to identify interesting patterns by analysis of data [2].

Table 1: Retrieval and Mining techniques Classification [2]

Purpose	Sources		
	Data	**Textual Data**	**Web Data**
Retrieving well-known facts or documents efficiently and effectively	Data Retrieval	Information Retrieval	Web Retrieval
Finding new patterns or knowledge previously unknown	Data Mining	Text Mining	Web Mining

From table-1 it is concluded that web mining research is the juncture of different areas (i.e. data retrieval, information retrieval, web retrieval, data mining and text mining). This paper is organized as follows. In section 2 Web mining is classified as content mining, structure mining and usage mining. In section 3 web mining is highlighted as trending research area for various research communities such as Databases, Artificial intelligence, Information retrieval and E-commerce. In section 4, Natural language processing technology is highlighted to explain how to handle unstructured data over the web using various NLP techniques (i.e. Part-of-Speech tagging). Finally, the paper is concluded in section 5.

2. Classification of Web Mining

Web mining can be categorized as content mining, structure mining and usage mining. Web content mining is a technique of fetching information from texts, images and other contents. Web structure mining is a technique of extracting information from linkages of web pages. Web usage mining is a process of take out information about the usage of web sites [7].

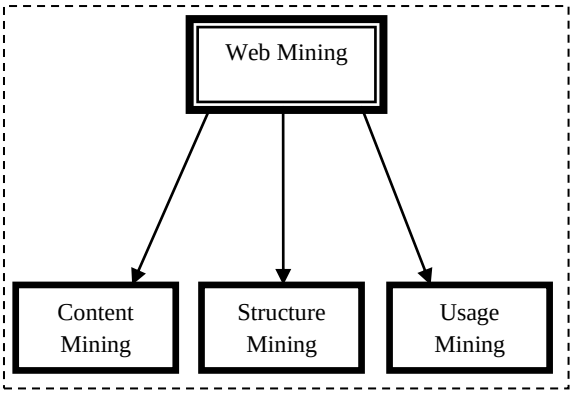

Figure 1: Classification of Web mining

Web Content Mining: Content mining is a process of finding information from millions of sources across the World Wide Web and mining these web data contents. Web data contents can be structured (i.e. data stored in the tables or HTML pages generated from database), semi- structured (i.e. HTML documents) or unstructured (i.e. text data) [11] [14].

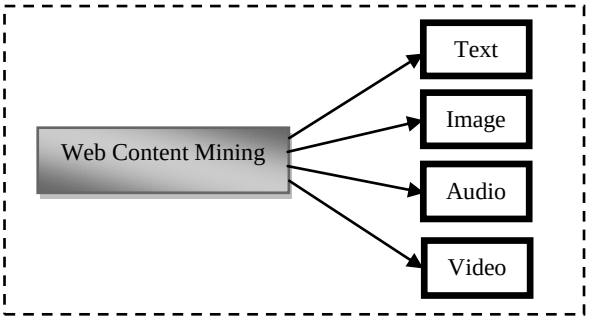

Figure 2: Taxonomy of web content mining

The un-structured properties of web data potency the web content mining in the direction of a further complex approach [3].

- Mining by developing a knowledge-base repository of the domain
- Interpretation of Mined Knowledge
- Iterative refinement of user queries for personalized search
- Process of Iterative Query Refinement

Web Structure Mining: As web content mining chiefly emphases the construction of inside the document, whereas web structure mining aim is to determine the link structure of the hyperlinks in the inter-document level [3]. It creates the structural framework about the web site along with web page. The structural outline consists of the following information [12]:

- Measure the frequency of the local links in the web tuples in a web table.
- Measure the frequency of web tuples in a web table containing links that are interior and the links that are within the same document.
- Measure the frequency of web tuples in a web table that having links that is global and the links that span different web sites [13].
- Measure the frequency of identical web tuples that appear in the web table or among the web tables.

The configuration of a web (directed graph) consists of web pages as nodes and hyperlinks as edges i.e. connection among related pages. Web structure graph terminology is as follows as given in table 2:

Table 2: Lexicon of web structure graph

Web-graph	A directed graph that exemplifies the web.
Node	Each Node represents the web page of the web-graph.
Link	Each hyperlink represents the directed edge of the web-graph

Figure 3 depicts the web graph structure by means of document structure and hyperlinks.

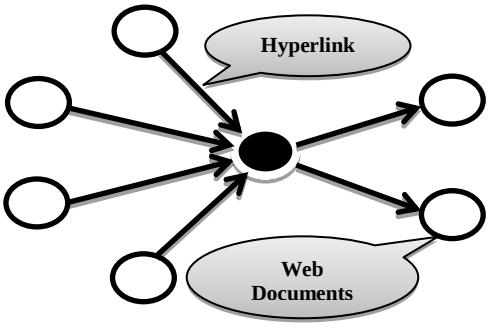

Figure 3: Web Graph Structure

A hyperlink connects a web page to a poles apart location within the same web page is called an intra-document hyperlink. A hyperlink that links two diverse pages is called an inter-document hyperlink [1].

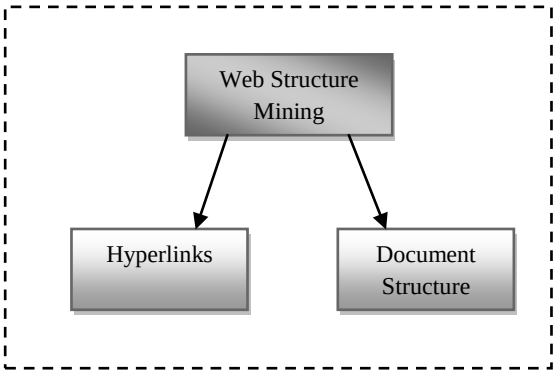

Figure 4: Web structure mining taxonomy

Web Usage Mining: Web usage mining is an activity that automatically discovers the user access [15] patterns from different web servers. It keeps track of earlier retrieved pages by a user that helps in identifying the distinctive behavior of the user and to make forecast about preferred pages [4]. Web usage mining includes three major phases i.e. preprocessing, pattern discovery and pattern analysis as shown in figure 5.

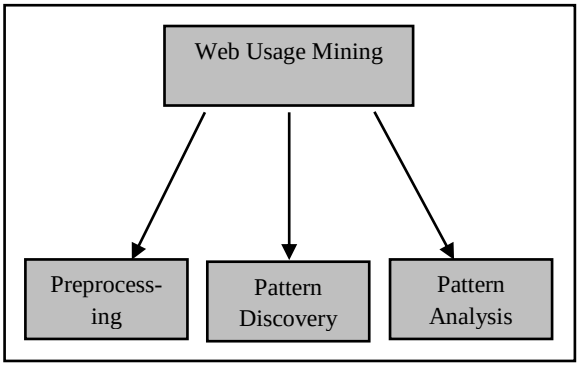

Figure 5: Segments of web usage mining

Web usage mining phases are defined as follows: -

- **Preprocessing:** This phase converts the raw usage data into the data abstractions. It includes usage preprocessing, content preprocessing and structure preprocessing. Preprocessing stage follow some steps such as data cleaning, efficient user identification, session identification and path completion, and transaction identification [6].

- **Pattern Discovery:** This phase includes different techniques such as association rules, clustering etc for pattern discovery. It draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition [10].

- **Pattern Analysis:** The purpose of pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. It requires the analysis of the structure of hyperlinks and the contents of the pages.

The difficulties in web usage mining occur due to the anomalies in existing data.

3. Applications of Web mining

Web mining spreads analysis much further by combining other corporate information with Web traffic data. Practical applications of Web mining technology are abundant, and are by no means the limit to this technology. Web mining tools can be extended and programmed to answer almost any question. It can be applied in following areas:

- **Managerial decision making:** Web mining can provide companies managerial insight into visitor profiles that helps in taking strategic actions by top management [4].
- **Marketing effectiveness:** Companies can have some subjective measurements through web mining regarding the effectiveness of their marketing campaign or marketing research, which will help the business to improve and align their marketing strategies timely.
- **Business related decisions:** In the business world, structure mining can be quite useful in determining the connection between two or more business web sites [7].
- **Accounting and Inventory:** Web mining also allows accounting, customer profile, inventory, and demographic information to be correlated with web browsing.
- **Improvement feedback:** Companies can identify the strength and weakness of their web marketing campaign through feedback from web mining, and can make the strategic adjustments accordingly [8].
- **Searching enhancement:** Search engine such as Google provides advanced and efficient searching capabilities by the usage of web mining [5]
- **E-commerce (Infrastructure):** Generate user profiles, targeted advertizing**,** fraud and similar image retrieval [9].
- **Information retrieval (Search) on the Web:** Automated generation of topic hierarchies**,** web knowledge bases**,** extraction of schema for XML documents.
- **Network Management:** results in performance management and fault management.

4. Handling of Web Content using Natural Language Processing

Web data consists of various types like structured data (i.e. data retrieved from backend databases), Semi-structured data (i.e. data organized as a hierarchy of blocks) and unstructured data (i.e. natural language text) as shown in figure 6.

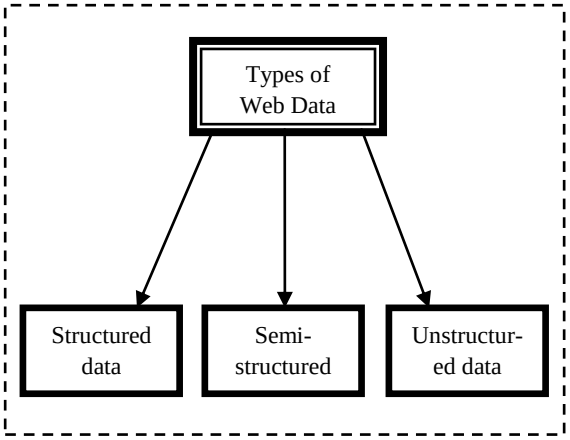

Figure 6: Types of Web data

The aforementioned web contents can be handled efficiently by following the below given techniques that is related to natural language processing to some extent. Natural language processing helps to understand how to manage unstructured data over the machine processing platform using various NLP techniques along with the web content mining.

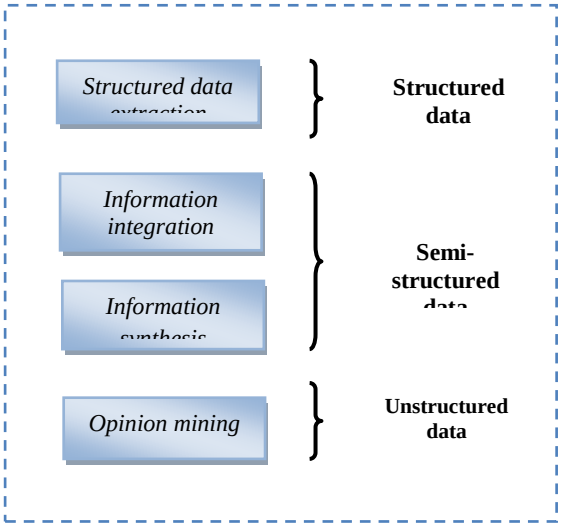

Figure 7: NLP techniques to handle different types of web content

Structured Data Extraction: web data or information is arranged/ managed as structured data objects on regular basis. Such as retrieval of various data records from databases. Extraction of these structured objects is possible by either the Wrapper induction technique (i.e. Supervised) or Automatic extraction (i.e. unsupervised) technique [16].

- *Wrapper induction technique* is based on machine learning.
- *Automatic extraction technique* is based on POS tagging (Part-Of-Speech). POS tagging (NLP technique) is to automatically assign part-of-speech tags (i.e. noun, verb, adjective etc.) to words in context [17].

Information Integration: Structured data objects extraction is followed by integration of data to design a consistent and reliable database. Integration can be in terms of schema match or data instance match.

- *Schema match:* From various data tables match the columns (e.g., Item names).
- *Data instance match:* From various data fields match the values, e.g., "Coke" = "Coca Cola"?

Information/ knowledge synthesis: It works upon a web search paradigm where a request for some words is given and a ranked list of pages is returned by search engine and top-ranked pages read by the user to find required information. This technique is adequate/suitable for navigational queries (i.e. specific information) but not for informational queries (i.e. open-ended research problems).

Opinion mining: Web content mining study is to extract precise sorts of information from text in Web documents e.g. Opinion (positive or negative) and factual information (i.e. Find economic data from rumors of different countries). Opinion mining or sentiment analysis purpose is to excerpt and abridge opinions.

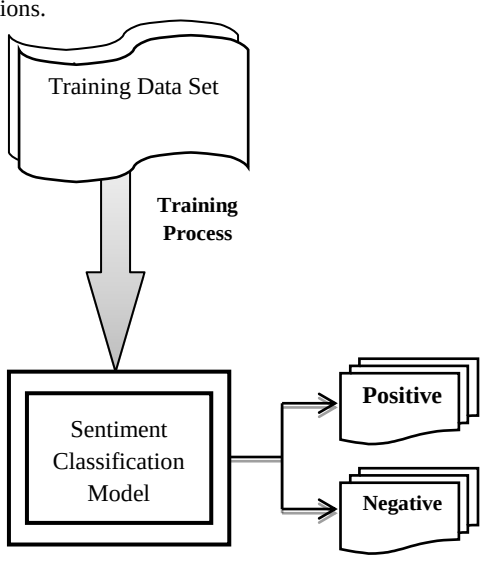

Figure 8: Opinion Mining

5. Conclusion

The increasing demand of the web has greatly evolved the web mining technology. Web mining is concerned with the mining of data from the various web documents, hyperlinks between documents and usage logs of web sites. Mining approach is to be followed can be either process-centric or data centric. This paper presents an overview of web content mining, web structure mining and web usage mining. Unstructured data of web contents can be handled by using natural language processing mechanism. An important research area in web mining is web usage mining which focuses on the sighting of interesting outlines in the glancing and steering data of web users. It helps in personalization of web content by having track of earlier retrieved pages by a user that result in identification of the distinctive behavior of the user and to make forecast about favorite pages. The outcomes formed by web usage mining can be exploited to expand the performance of web servers and web-based applications.

6. References

[1] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, "Web Mining: Concepts, Applications, and Research Directions".

[2] Hsinchun Chen and Michael Chau, "Web Mining: Machine learning for Web Applications", Annual Review of Information Science and Technology.

[3] Sarita Dalmia, "Web Mining : Survey and Research".

[4] Ankita Kusmakar, Sadhna Mishra, "Web Usage Mining: A Survey on Pattern Extraction from Web Logs", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 9, September 2013.

[5] Monika Yadav Mr. Pradeep Mittal, "Web Mining: An Introduction" , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013.

[6] DeMin Dong, "Exploration on Web Usage Mining and its Application", International Workshop on Intelligent Systems and Applications, Pp. 1.

[7] Robert Cooley, Bamshad Mobasher, Jaideep Srivastava , "Web Mining: information and Pattern Discovery on the WWW".

[8] Mary Garvin, "Data Mining and the Web: What They Can Do Together".

[9] R. Kosala, H. Blockeel, "Web Mining Research: A Survey", in SIGKDD Explorations, ACM, Volume 2, Issue 1, July 2000.

[10] J. Srivastava, R. Cooley, M.Deshpande, P-N. Tan. " Web Usage Mining: Discovery and Applications of usage patterns from Web Data", SIGKDD Explorations, Volume 1, Issue 2, 2000

[11] Etzioni, "The World Wide Web: Quagmire or gold mine" , Communications of the ACM, Vol. 39, issue ll, 1996, pp. 65-68.

[12] Cooley, R., Mobasher, B., & Srivastava, "Web mining: information and pattern discovery on the World Wide Web" , In Proceedings of the 9th ZEEE International Conference on Tools with Artificial Intelligence, 1997, pp. 558-567.

[13] Liu Bin, "Web Data Ming Exploring Hyperlinks, Contents, and Usage Data".

[14] Kshitija Pol, Nita Patil, Shreya Patankar, Chhaya Das, " A Survey on Web Content Mining and extraction of Structured and Semistructured data".

[15] Bing Liu, "From Web Content Mining to Natural Language Processing", 2007.

[16] M. Rajman, R. Bseancon, "Text Mining: Natural Language Techniques and Text Mining Applications"

[17] Lihui Chen, Wai Lian Chue, "Using Web structure and summarisation techniques for Web content mining".