A Review on Basics in Web Mining

L.Shobby Neeta Fancy[1], A.Rajamurugan[2]

PG Student, Information Technology, Anna University Regional Centre, Coimbatore, India [1]

Teaching Fellow, Information Technology, Anna University Regional Centre, Coimbatore, India [2]

Abstract*:* Today World Wide Web (WWW) is growing into eternity and it has massive wealth of information. Ordinary people also use the web to retrieve the needed and related semantic information. So that, mining the web is very essential in order to retrieve the necessary information for any user. Web mining can be categorised into the following, namely, Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM). These categories can be reviewed by this paper. The two link analysis algorithms namely PageRank and Hyperlink Induced Topic Search are also discussed in the Web Structure Mining.

Keywords*:* Web mining, Web content, Web structure, Web usage, PageRank, Hyperlink Induced Topic Search**.**

I. INTRODUCTION

Due to massive collection of information in the web, retrieving the semantic and needed information from the web has become more critical. For the effective retrieval of information from the web, there is a high need to apply some effective techniques over the web. So that, applying data mining techniques over the web to automatically extract meaningful patterns from the web documents becomes essential. Data mining techniques can be used to make the web more intractable and profitable too. This paper is the detailed review of different web mining techniques.

II. WEB MINING

Web mining is the application of data mining techniques to automatically discover new patterns. The discovered patterns may be cluster of similar documents, classification of new documents or frequent item set discovery that has been used to relate pages being referred together in a single session.

A. Tasks Involved In Web Mining

Web mining can be divided into the following tasks[1],[2],[3].

- Resource finding – The task of extracting relevant information resources from the web.
- Information selection and pre-processing – It means selecting and pre-processing of specific and intended information from the retrieved web resources.
- Generalization – This is the process of automatically discovering general patterns within the individual websites or even from across multiple websites.
- Analysis – It generally refers to the evaluation, validation and interpretation of the mined patterns.

B. Steps Involved In Web Mining

The mainly involved steps in web mining are listed below [4].

- Fetching the content from the web.
- Parsing the data.
- Analysing the data like tokenizing, rating and classifying.
- Producing useful data or patterns from the web.

Fig.1. depicts these steps.

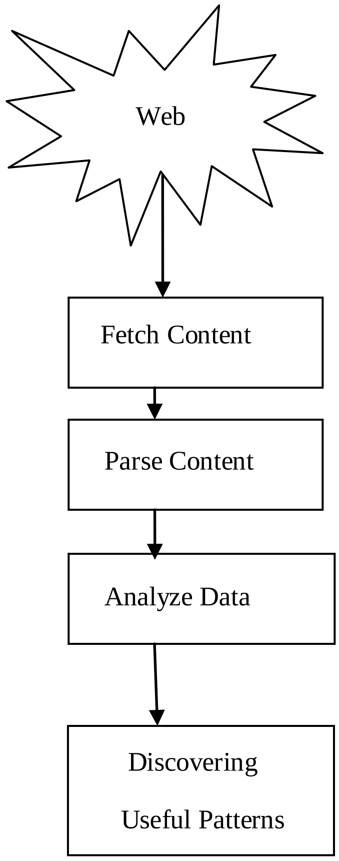

Fig. 1. Steps of web mining

III. CLASSIFICATION OF WEB MINING

Mining from web can be classified into three types [1].

- Web content mining
- Web structure mining
- Web usage mining

Fig.2. illustrates this classification.

A. *Web Content Mining(WCM)*

Applying data mining techniques to different contents namely unstructured document (text) or semi-structured document (html) or structured document (data from row-column database) to retrieve the needed and semantic information from the web is called as Web Content Mining (WCM).

As mentioned earlier web content can be broadly classified into semi-structured, unstructured or structured data which can be extracted from databases.

Structured data means the data that has been located on any fixed fields in a file or record. Structured data can be presented in any web page either in the form of data that has been retrieved from the database table or like registration forms with pre-defined fields. In these types of data, the fields of a database are pre-defined. Structured data is very easy and useful for querying and retrieving the semantic information from the web. (e.g.) The data will be retrieved from relational database.

Semi-structured data does not conform to any formal structure of relational databases or other data tables. This contain tags to separate meaningful information and insist hierarchies of records and fields within the data. Semi-structured data is also referred as self-describing structured data.(e.g.) html pages.

Unstructured data means the data that has not been located in the row-column databases. The text documents and the multimedia contents are the unstructured data files. E-mail messages, videos, photos and audios are the best examples of unstructured data.

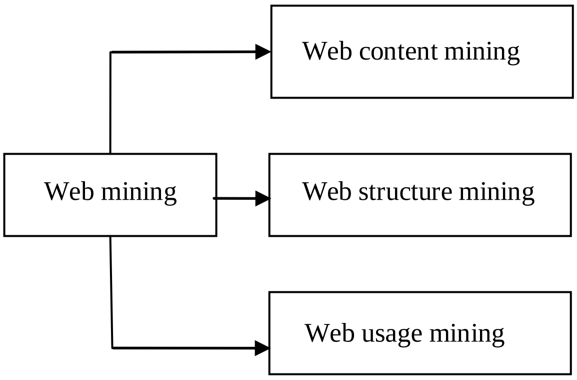

Fig. 2. Classification of web mining

Normally a web page consists of text, graphics, tables, data blocks and data records. Intelligent software agents are also called as softboats which are used to find and filter the available information from the web with respect to the query issued by the user. WCM is very helpful to screen the specific data from the web. This is used in Information Retrieval (IR)

field. WCM is the scanning and mining of text, pictures, graphics of a web page to determine the content's relevance to match the user's query. This web content mining has been completed after the clustering of web pages through the structure mining process and provides the results based upon the level of relevance to the issued query. The results are arranged in a logical sequence from highest level to lowest level in terms of relevance.

1) Mining of Text in Information Retrieval [5]:

Data which are applied to data mining methods must be in the format of spreadsheet. But the data that has been presented in the web will be in the document format too. So the initial step of text mining is to convert the data in the document format into the data belonging to the spreadsheet format. For this, the input is the collection of documents and the output is the binary data that should be presented in the spreadsheet format. In this format, the row represents the collection of documents and the column represents the collection of words. These collections of words are called as dictionary. In the output, 1 represents the presence of a word and 0 represents the absence of a word. Spreadsheet representation is also called as row-column representation or matrix format representation. Spreadsheet behaves like a reasonable conceptual model of data. All the values in the text mining spreadsheets are positive. But in data mining positive and negative values of the attributes have been considered.

TABLE I

A BINARY SPREADSHEET OF WORDS IN DOCUMENTS

	Web (Word1)	HTML (Word2)	Crawler (Word3)	Ranking (Word4)
Doc1	1	0	1	0
Doc2	1	1	0	1
Doc3	1	1	1	1

In IR, a query is presented to the search engine. The whole query is treated like a new document. For this new document, the similarity will be measured with other stored documents. After that, the matched documents will be retrieved as the responses of the search engine. Similarity is measured using shared word count, word count and bonus or cosine similarity. The spreadsheet model has been used for this task. The user's query (i.e.) the new document is treated as a new row in the spreadsheet. This new row will be compared to all other rows and the most similar rows with their related documents are retrieved as the responses to the user's query.

In TABLE I, row represents documents and column represents words. For example the word HTML can be available in document 2 and document 3 but does not present in the document 1.

The commonly used web content mining tools are Screen Scraper(SS) , Web Info Extractor(WIE), Mozenda and Web Content Extractor(WCE).

B. *Web Structure Mining (WSM)*

Web structure mining is based on the link analysis and topology of the web graph. Web graph is the collection of web pages where each page is represented as a node in the graph and the nodes are interconnected via edges. Here edges represent the hyperlinks that are the interconnection between the web pages. Web graph is the edge weighted, directed graph. WSM is mainly used in Information Retrieval (IR) to rank the responses from the web by analysing the hyperlink. Link based algorithms are generally used for WSM.

The measure of merit of a web page is given below:

- Web pages with large in-degree indicate the power of authority.
- Web pages with large out-degree indicate the large coverage.
- Similarity to a driving query.

1) Link Based Algorithms:

In WSM, two mainly used link based algorithms are given below [6].

- PageRank(PR)
- Hyperlink Induced Topic Search(HITS)

Both are mainly used to rank the responses retrieved from the search engine.

PageRank (PR) - Google uses PR algorithm to rank the retrieved Search Engine Result Pages (SERPs). In this, a web page depends on number of links pointing to it and their ranks.

Fig.3. shows the links between pages.

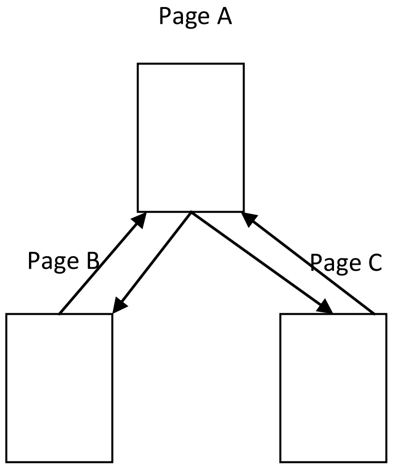

Fig.3. Links between pages

To find the PR of a page the following formula can be used.

PR (A) =d+ (1-d)*$\sum_j$(PR(T_j)/C (T_j))

Here, 'A' represents web page for which the PR has to be calculated.

'd' represents a minimum value being assigned to any web page.

'C (T)' represents number of outgoing links from the page 'T'.

Hyperlink Induced Topic Search (HITS) - In Hyperlink Induced Topic Search (HITS), a query dependent web graph is chosen for analysis.

According to HITS, the web includes two popular pages.

- "Authorities" means to which many pages link.
- "Hubs" - which are set of links to authorities.

Every web page 'u' behaves like both hubs and authorities with their corresponding scores named as hub score h[u] and authority score a[u].

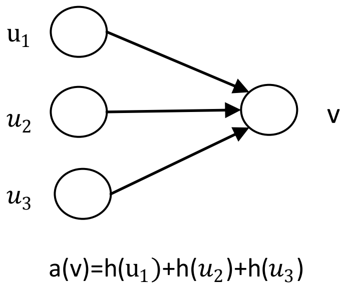

Fig. 4. Authority score

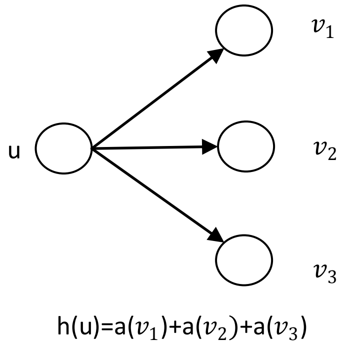

Fig. 5. Hub score

Fig.4. and Fig.5. represent the calculation of authority and hub score.

The major steps involved in HITS are [6],

- Send a query to the IR system and obtain the web pages which have been matched to the query (i.e.) nodes of the graph are called as root set 'R'.
- For any node 'u', that is the neighbour of any r€R via an incoming or outgoing edge (i.e.) (u,r) or (r,u) is also included for that query. This is called as an expanded set.
- Root set + Expanded set= Base set
- Power iterations will be run on the hub and authority scores together.
- The top ranking authorities and hubs will be reported.

PageRank of any web page is computed independent of the user's query. It can be computed before the query is issued. But in HITS, the web graph is chosen depending on the user's query. This measure should been computed during the issuing of the query. This is the major difference between PageRank and Hyperlink Induced Topic Search algorithms.

C. *Web Usage Mining(WUM)*

Web usage mining is one of the types of web mining. In this, the data mining techniques can be applied to explore user access and navigational patterns from the web. This is called as usage patterns. The WUM can be operated on the server logs to discover the usage patterns from the web. The web usage data includes data from web server access logs, browser logs, user profiles, user sessions, cookies, queries that has been issued by the user, amount of time spent by the user in a specific web page and click through URLs. These data depicts the behaviour of the user.

The server log is a log file which has been created automatically by a server. The contents of the log file are given below [7]:

- Client IP Address
- User ID
- Date/Time
- Page requested
- HTTP code
- Bytes served
- User agent
- Referrer

1) Main Phases of WUM [8]:

- Pre-processing – The usage, content, structure information available in the data sources are converted into the data abstractions.
- Pattern discovery – The pre-processing data abstractions can be used here. Distinctive methods and algorithms are applied here to extract the usage patterns.
- Pattern analysis – This is the last step in WUM process. The major goal of pattern analysis is to filter out uninteresting rules or patterns found in the pattern discovery phase.

2) Application Areas of WUM [8]:

- Business intelligence
- Personalization
- Site modification
- System improvement
- Usage characterizations
- Automatic recommendation systems

In business intelligence web usage mining can be mainly used by the website administrators who are all involved in e-commerce and marketing. Through this, the marketing advertisers suggest some more related recommendations to the web user who is involved in the online shopping transaction to increase their profit and to advertise their products. In personalization WUM is very helpful in web personalization. Personalizing web means, for a given query, the web search engine produces different SERPs or reorganizes the SERPs differently for different users. For this the intuition of the user is captured by the usage patterns. In site modification WUM provides feedback relevant to the user's preferences to the site administrators. Using this information, the website administrators can redesign the structure and content of the website based upon the usage patterns.In system improvement WUM is very helpful to understand the web traffic behaviour. This can be useful to develop policies for web caching, load balancing and data distribution. In web services, security is the major issue. For that, WUM provides the usage patterns which are very useful to detect the fraudulent activities and intruders.

IV. CONCLUSION

This paper discusses about the web mining concepts and categories of web mining. The categories of web mining are Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM). In this paper each category has been explained separately. Web content mining is used to retrieve the semantic information from the unstructured text document. In web structure mining, the link analysis algorithms namely PR and HITS are used to rank the responses of the search engine relevant to the user's query. In web usage mining the usage patterns are extracted from the user's behavior.

REFERENCES

[1] Raymond Kosala and Hendrik Blockeel," Web Mining Research: A Survey", SIGKDD Explorations, vol.2, issue 1.

[2] Chintandeep Kaur and Rinkle Rani Aggarwal," Web Mining Tasks and Types: A Survey", IJRIM, vol.2, issue 2.

[3] Renu Sharma,"A Framework to Compare Web Mining Types", International Journal of Advanced Research in Computer Science and Software Engineering,vol.3,issue 7,July 2013.

[4] Ken Krugler,"Web Mining in the Cloud",2009.

[5] Sholom M.Weiss, Nitin Indurkhya, Tong Zhang and Fred J.Damerau,"Text Mining-Predictive Methods for Analysing Unstructured Information", Springer, 2005.

[6] Soumen Chakrabarti,"Mining The Web-Discovering Knowledge from Hypertext Data", Morgan Kaufmann Publishers, An imprint of Elsevier Science, 2003.

[7] Robert Walker Cooley,"Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data", Thesis for the degree of Doctor of Philosophy, The University of Minnesota.

[8] Jaideep Srivastava, Robert Cooley, Mukund Deshpande and Pang-Ning Tan,"Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, vol.1, issue 2.

BIOGRAPHIES

L.SHOBBY NEETA FANCY has obtained B.TECH. Information Technology in University College of Engineering Tindivanam (A Constituent College of Anna University Chennai) and doing M.TECH. Information Technology in Anna University Regional Centre Coimbatore.

Mr.A.RAJAMURUGAN has obtained M.TECH. Mainframe Technology in Anna University Regional Centre Coimbatore and he is working as a teaching faculty in Department of Information Technology in the same regional centre.