Volume 7, Issue 5, May 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Web Pattern Analysis Using Web Structure Mining

S Sundeep Kumar
M.Tech Student, Biet Lucknow,
Uttar Pradesh, India

Mahesh Kumar Singh
Asst. Professor, Biet Lucknow,
Uttar Pradesh, India

DOI: 10.23956/ijarcsse/SV7I5/0274

Abstract: ***The World-Wide-Web contains a large amount of information. Everyone can store and retrieve the information from Web. It is difficult to find the relevant piece of information from Web. Extracting the important information from Web is called Web Mining. Web mining is the application of data mining techniques to extract knowledge from Web data, including Web documents, hyperlinks between documents, usage logs of Web sites, etc. Web mining technologies are best suited for Web information extraction and information retrieval. Web mining is one of the mining technologies, which applies data mining techniques in large amount of Web data to improve the Web services. We are going to give a brief description of Web mining and its categorization namely: Web content mining, Web usage mining and Web structure mining.***

Keywords: ***Natural Language Processing (NLP), Intrapage Structure.***

I. INTRODUCTION

With the rapid increase in the use of Internet, and the amount of data that now available on the Web, it has become very important to determine which data is relevant and which is irrelevant. Web search has become amazingly powerful in its ability to discover and exploit nearly any kind of information within the billions of pages that comprise the Web. Currently, almost any search engine faces the increasingly difficult challenge of collecting, storing, processing, retrieving and distributing Web data for users with different search intentions, needs and backgrounds. While traditional algorithmic search engines have been very successful in dealing with relatively simple keywords based Web search, recently there has been a surge of interest in exploring new territories of the Web due to the appearance of new user groups and search needs requiring the development of novel Web search applications.

II. WEB MINING

Web mining is the application of data mining techniques to extract useful information and knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc. to improve the web services. Web mining refers to the overall process of discovering potentially useful and previously unknown information or knowledge from the Web data. A natural combination of Data Mining and World Wide Web may be referred to as Web Mining.

Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents. It consists of following tasks:

- *Resource finding:* It involves the task of retrieving intended web documents. It is the process by which we extract the data either from online or offline text resources available on web. It includes information retrieval and extraction from web pages.
- .*Information selection and pre-processing*: It involves the automatic selection and pre processing of specific information from retrieved web resources. This process transforms the original retrieved data into information. Making web data suitable for mining is preprocessing.
- *Generalization*: It automatically discovers general patterns at individual web sites as well as across multiple sites. Data Mining techniques and machine learning are used in generalization. Result displayed in a web search is aggregation of multiple web documents.
- *Analysis*: It involves the validation and interpretation of the mined patterns. It plays an important role in pattern mining. A human plays an important role in information on knowledge discovery process on web.

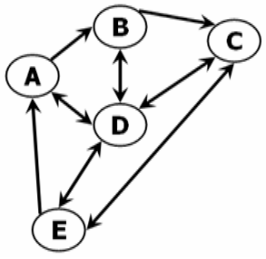


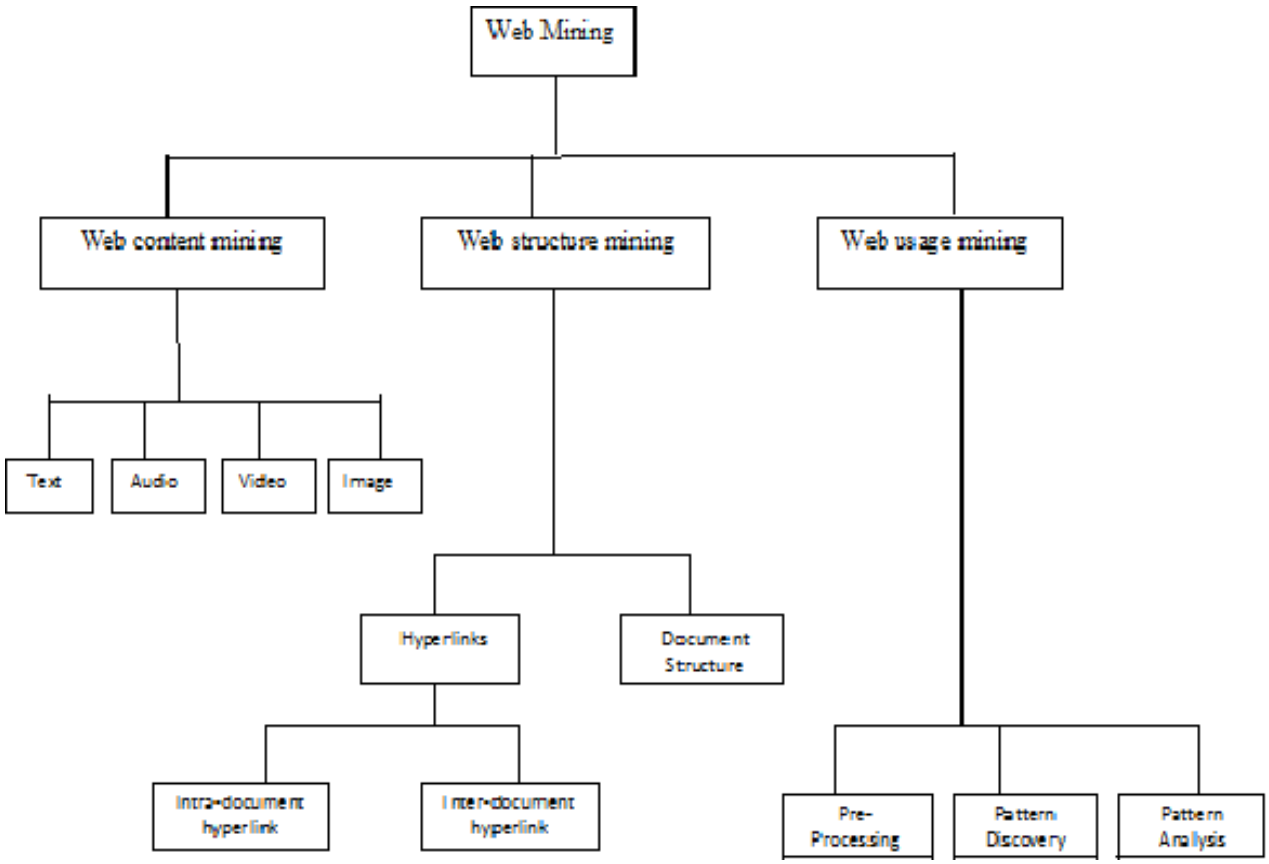
Figure (1) Website structure

Website structure can be easily understood from figure (2). As we know that website is a collection of related web pages containing images, videos or other digital assets? In figure (1) A, B, C, D, E is different pages of website. It is clear that if hyperlinks are available then we can easily move between pages. In Web mining website structure is also important. If the website structure is complex then it will take less time to move between pages. It is also clear from the fig(2) that if there are so many links (known as hyperlink) pointing to one another then the structure of the website will get complex. So much number of hyperlinks will generate a complex website.

WEB MINING CATEGORIZATION

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined as shown in Figure-

1. Web content mining,
2. Web structure mining
3. Web usage mining.



WEB MINING vs. DATA MINING

Web Mining is the application of data mining techniques to discover and retrieve useful information and patterns from the World Wide Web documents and services.

There are general classes of information that can be discovered by web mining:

Content – data from Web documents – text & graphics

Structure – data from Web Structure – HTML or XML tags

Usage – data from Web log data – IP addresses, date & time access

User Profile – data that is user specific – registration and customer profile

When comparing web mining with traditional data mining, there are three main differences to consider:

Scale: In traditional data mining, processing 1 million records from a database would be large job. In web mining, even 10 million pages wouldn't be a big number.

Access: When doing data mining of corporate information, the data is private and often requires access rights to read. For web mining, the data is public and rarely requires access rights. But web mining has additional constraints, due to the implicit agreement with webmasters regarding automated (non-user) access to this data. This implicit agreement is that a webmaster allows crawlers access to useful data on the website, and in return the crawler (a) promises not to overload the site, and (b) has the potential to drive more traffic to the website once the search index is published. With web mining, there often is no such index, which means the crawler has to be extra careful / polite during the crawling process, to avoid causing any problems for the webmaster.

Structure: A traditional data mining task gets information from a database, which provides some level of explicit structure. A typical web mining task is processing unstructured or semi-structured data from web pages. Even when the underlying information for web pages comes from a database, this often is obscured by HTML markup.

There is misconception in people that Web mining is different from Data mining. Now it is clear from above explanation that web mining use data mining techniques in Web data. Data mining is the collection of (Anomaly detection, Association rule learning, Clustering Classification, Regression, Summarization) activities, while web mining uses these activities to extract the information from web data.

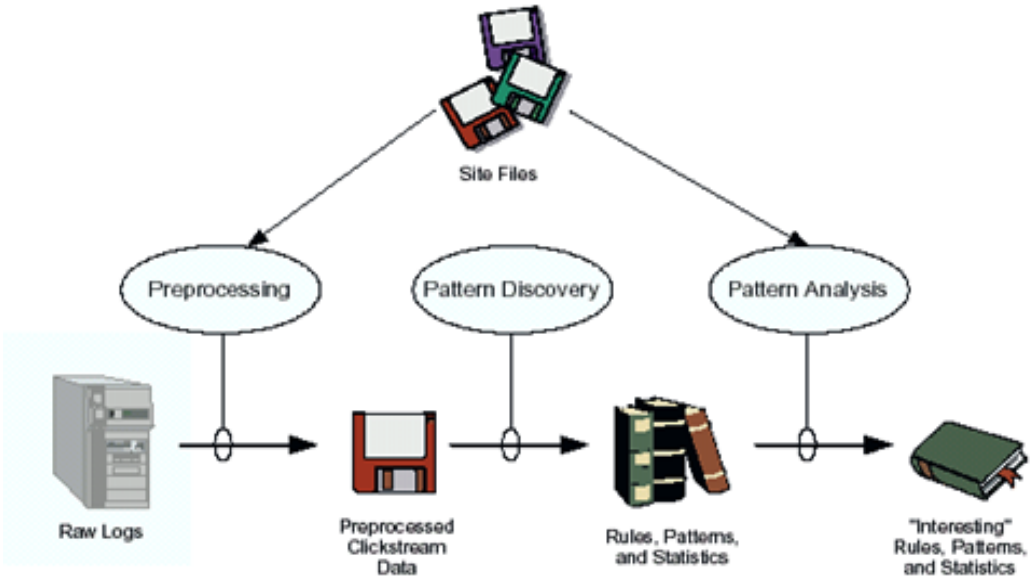


Figure (2) Web Usage Mining Process

Web Mining Process tasks are already explained in introduction chapter. Here it is shown diagrammatically. All the phase namely are Resource finding, Information selection and pre-processing, Generalization and analysis. Fig(4) is representing Web Usage Mining Process.

Web structure mining is one of the categorization of Web mining, as explained previous chapter and shown in figure given below.

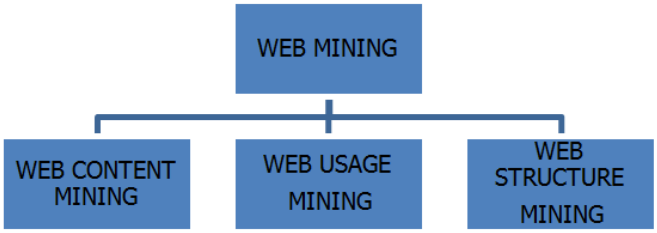


Figure (3)

WEB STRUCTURE MINING

Web Structure Mining is the process of discovering structure information from the Web. Web structure mining aims to generate structural summary about web sites and web pages. The focus of structure mining is therefore on link information, which is an important aspect of web data. Given a collection of interconnected web documents, interesting and informative facts describing their connectivity in the web subset can be discovered.

The following structural information can be generated from the web tuples stored in the web tables .

Measuring the frequency of the local links in the web tuples in a web table. Local links connect the different web documents residing in the same server. This informs about the web tuples (connected documents) in the web table that have more information about inter-related documents existing at the same server. This also measures the completeness of the web sites in a sense that most of the closely related information are available at the same site. For example, an airline's home page will have more local links connecting the "routing information with air-fares and schedules" than external links.

Measuring the frequency of web tuples in a web table containing links which are interior; links which are within the same document. This measures a web document's ability to cross-reference other related web pages within the same document. This also measures the flow of the web documents. For example, a news-paper should always refer to other news items locally (within the same news-paper).This information depicts that the relevant information is available within the same file.

Measuring the frequency of web tuples in a web table that contains links that are global; links which span different web sites. This measures the visibility of the web documents and ability to relate similar or related documents across different sites. For example, research documents related to "semistructured data" will be available at many sites and such sites should be visible to other related sites by providing cross references by the popular phrases such as "more related links". Also, in case of a document like a research paper, it should have more external links as it should refer to other related papers. This expresses a research paper's ability to cross-reference other related work.

Measuring the frequency of identical web tuples that appear in a web table or among the web tables. This measures the replication of web documents across the web warehouse and may help in identifying, for example, the mirrored sites. This information concludes that some web pages provide integrated information on various topics. We have also used duplicate web tuples to identify, for example, visible web pages etc.

Another interesting issue is to discover the nature of the hierarchy or network of hyperlinks in the websites of a particular domain. For example, with respect URLs with domains like .edu, one would like to know how most of the web sites are designed with respect to information flow in educational institutes. What is the flow of the information they provide and how are they related conceptually. Is it possible to extract conceptual hierarchical information for designing web sites of a particular domain. This may help in generalizing the flow of information in web sites representing information in some particular domain. This will help for example in building a common web schema or wrappers

If a web page is directly linked to another web page or are near to each other then we would like to discover the relationships among those web pages. These relationships might be of the following types. The two web pages might be related by synonyms or ontology or having similar topics, both the web pages are in the same server and in that case both the pages may be authored by the same person.

What is the in-degree and out-degree of each node (web document)? What is the meaning of high and low in- and out-degrees? For example, a high in-degree may be a sign of a very popular web site or document. Similarly, a high out-degree may be a sign of luminous web site. Out-degree also measures a site's connectivity.

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages. Web Structure Mining can be regarded as the process of discovering structure information from the Web. This type of mining can be further divided into two kinds based on the kind of structural data used.

III. TYPES OF WEB STRUCTURE MINING

Web structure mining can be further divided into three categories based on the kind of structured data used.

- Web graph mining
- Web information extraction
- Deep Web mining

Web graph mining

Compared to a traditional document set in which documents are independent, the Web provides additional information about how different documents are connected to each other via hyperlinks. The Web can be viewed as a (directed) graph whose nodes are the Web pages and whose edges are the hyperlinks between them. There has been a significant body of work on analyzing the properties of the Web graph and mining useful structures from it (Page et al., Kleinberg, Bharat & Hen zinger, Gibson, Kleinberg). Because the Web graph structure is across multiple Web pages, it is also called *interpage structure.*

In terms of web graph mining, World Wide Web is considered as graph. The study of the web as a graph is not only fascinating in its own right, but also yields valuable insight into web algorithms for crawling, searching and community discovery, and the sociological phenomena which characterize its evolution.

Web information extraction (Web IE):

In addition, although the documents in a traditional information retrieval setting are treated as plain texts with no or few structures, the content within a Web page does have inherent structures based on the various HTML and XML tags within the page. While Web content mining pays more attention to the content of Web pages, Web information extraction has focused on automatically extracting structures with various accuracy and granularity out of Web pages. Web content structure is a kind of structure embedded in a single Web page and is also called *intrapage structure*. Extracting structured data from web pages is a key challenge due to the presence of noise. Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases this activity concerns processing human language texts by means of natural language processing (NLP). Recent activities in multimedia document processing like automatic annotation and content extraction out of images/audio/video could be seen as information extraction.

Deep Web mining:

Besides Web pages that are accessible or crawlable by following the hyperlinks, the Web also contains a vast amount of non crawlable content. This hidden part of the Web, referred to as the *deep Web* or the *hidden Web* (Florescu, Levy, & Mendelzon, 1998), comprises a large number of online Web databases. Compared to the static surface Web, the deep Web contains a much larger amount of high-quality structured information (Chang, He, Li, & Zhang). Automatically discovering the structures of Web databases and matching semantically related attributes between them is critical to understanding the structures and semantics of the deep Web sites and to facilitating advanced search and other applications. At this time crawlers cannot effectively query online databases, such data are invisible to search engines, and thus

The Deep Web remains largely hidden from users. To enable effective access to databases on the web, since April 2002, some systems have been presented.

IV. PARTITIONING ALGORITHM

Because we know that page ranking algorithms work in hyperlinked environment. Its best suitable way to analyze pages in hyperlinked space, hence in proposed algorithm we first read the web page and then represent the web page as a vector in total hyperlinked space but as we know that it can form very large dimensional vector because of large number of websites exists. Hence to reduce the size and complexity we only consider the first section of the url (i.e. from url http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html we will only take http://home.dei.polimi.it) during first clustering loop and then re-cluster each cluster using full url. Now pages are grouped in to N cluster, where N is the no of processing units. But there is no guarantee that it will converge to the global optimum. So it is repeated with different starting conditions until each cluster come to almost similar size (less than 10% standard deviation).

The algorithm can be written in steps as follows:

1. Represent each page by their hyperlinks
2. Count total no. of base urls from all Pages (suppose U)
3. Represent each page by U-dimensional vector
4. Perform k-means clustering for N groups, where N is the no. of total processing units
5. Calculate the size of each cluster
6. If standard deviations of size of clusters are not less than 10% then repeat from the step 4 with different initial values.
7. Now Count total no. of urls in one cluster (suppose U1)
8. Represent each page by U1-dimensional vector
9. Perform k-means clustering for N groups, where N is the no. of total processing units
10. Calculate the size of each cluster
11. If standard deviations of size of clusters are not less than 10% then repeat from the step 9 with different initial values.
12. Calculate the distance among all cluster & place the clusters with minimum distance in same machine.

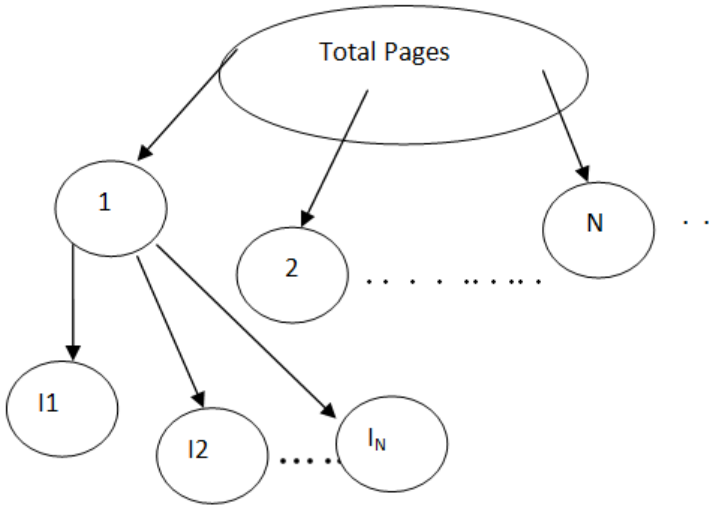


Figure (4) - Showing Clustering approach

V. CONCLUSION AND FUTURE SCOPE

Its been observed that using clustering an algorithm can be designed which is based on partitioning model. The algorithm shows that each partitioned cluster contains average of 80% of all referred pages refer by the pages within that cluster even the percentage remains approximately unchanged from the variations of cluster numbers. The processing time increases linearly as the no of processing node increases.

As the proposed technique is good for clustering, but the time consumed in clustering increases with increase in number of clusters. The reason behind is property of k-means clustering which is sensitive to initial condition. It causes the non uniform cluster information for which clustering is repeated. Hence this problem can be overcome by the use of k-means ++ clustering.

REFERENCES

[1] http://www.about.com

[2] Olston, C. and Chi, E. H. (2003) ScentTrails: Integrating Browsing and Searching on the Web. *ACM Transactions on Computer-Human Interaction (TOCHI)*, Vol. 10, No. 3, pp. 177-197.

[3] Perkowitz, M. and Etzioni, O. (1998) Adaptive Web sites: automatically synthesizing Web pages. In *Proc. of AAAI'98*, pp. 727-732, July 26-30, Madison, Wisconsin, USA, ISBN 0-262-51098-7, AAAI Press.

[4] Netcraft. Web server survey, 2004.

[5] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey, *SIGKDD Explorations, ACM SIGKD July 2000.*

[6] O.etzioni. The world wield web: Quagmire or Gold Mining. *Communicate of the ACM, (39)11:65-68, 1996.*

[7] Rekha Jain, Dr. G. N. Purohit, Page Ranking Algorithms for Web Mining, *International Journal of Computer Applications (0975 – 8887) Volume 13– No.5, January 2011.*

[8] Masashi Toyoda, Masaru Kitsuregawa What's Really New on the Web? Identifying New Pages from a Series of Unstable Web Snapshots, *WWW 2006, May 23–26, 2006, Edinburgh, Scotland. ACM 1-59593-323-9/06/0005.*

[9] Hong T, Chiang M, Wang S H, "Mining weighted browsing patterns with linguistic minimum supports", 2002 *IEEE International Conference on Systems, Man and Cybernetics,* 2002,Yasmine Hammamet, Tunisia, pp. 635-639.

