The International Journal Of Engineering And Science (IJES)
|| Volume || 5 || Issue || 5 || Pages || PP -27-31|| 2016 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Comparable Analysis of Web Mining Categories

[1]Anmol Kaur , [2]Dr.Raman Maini
[1]*M.Tech Student ,Department of Computer Engineering Punjabi University, Patiala , Patiala, India*
[2] *Professor Department of Computer Engineering Punjabi University, Patiala Patiala, India*

----------ABSTRACT----------
Web Data Mining is the current field of analysis which is a combination of two research area known as Data Mining and World Wide Web. Web Data Mining research associates with various research diversities like Database, Artificial Intelligence and Information redeem. The mining techniques are categorized into various categories namely Web Content Mining, Web Structure Mining and Web Usage Mining. In this work, analysis of mining techniques are done. From the analysis it has been concluded that Web Content Mining has unstructured or semi- structure view of data whereas Web Structure Mining have linked structure and Web Usage Mining mainly includes interaction.

Date of Submission: 18 April 2016 Date of Accepted: 05 May 2016

I. Inroduction

The knowledge can be spread through World Wide Web . Internet has become an important part of our today's busy schedule. It has turned out the manner of working business, education handling the organisation etc. The Web is large collection of information which is huge and dynamic in nature. That's why the complexity also increases to handle this abundant data. But the user satisfaction is must. Because the user want the perfect answer about the topic which he want to search. Different users have different needs and level of satisfaction according to their area or field. Like students want to examine the answer about the topics of study, business mind people like to analyse the customer's requirements. Everyone wants techniques to meet their needs. Mining can be implemented to find the Data Mining tools to find the required knowledge from internet. This gathered information is apply to obtain more command and observe the information make forecast, what would the right option and the fair appeal to move go ahead [3]. According to studies, Web Mining can be categorized into the following three types of categories namely Web Content Mining, Web Structure Mining, Web Usage Mining respectively. Web mining techniques are decomposed into the following subtasks:

1. **Resource Discovery:** it takes care to find the web information from various sources.
2. **Information selection and pre-processing**: the data which is collected from web is selected and pre-process it automatically.
3. **Generalization:** patterns are automatically discovered at both the various sites and discrete sites.
4. **Analysis:** it certify the data which is mined [7].

II. Different Categories of Web Data Mining

Description of different types of Web Data Mining namely Web Content Mining, Structure Mining and Web Usage Mining. This categorization is depicted in Figure 1.

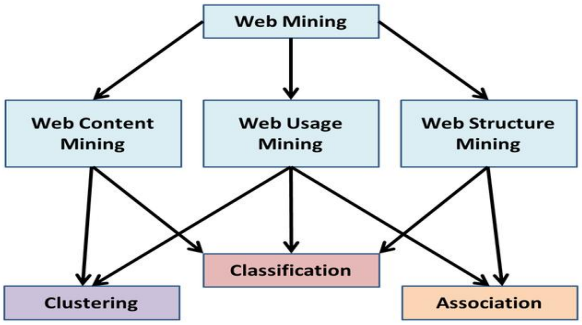

Fig 1. Web Mining Categorization [3]

2.1. Web Content Mining

The Web Content Mining explains the automatic exploration of information which is available on web. Content Mining deal with examining of topic, images and graphs to discover the applicable material. Many of them are semi-structured or some unstructured in nature. This examining is over when the collection is done through structure mining and produces result build upon the stage of applicability. With the bulky load present on the web this mining gives the results based upon the priority [2].The Web Content Mining is aimed toward particular data regulated by the client search information in the various search engines. This grants for the scanning of the whole Web to gather the chunk of content triggering the scanning of peculiar Web pages which are within those clusters. The outcome pages communicate with search engines with the order of highest level to lowest level. This mining allows to minimize the inappropriate data [2;3].

The Content mining is effectual or constructive when it is used while dealing with particular database. As for example online universities utilize a library system which helps in recalling articles relevant to their fields of study. This particular database allows to drag only those information related with the concerned subjects. The advantage of this mining is that it categorize, organize and provides the possible results available on the internet. Web mining helps to raise prolific uses of mining for business, web designing and search engines operations[2].

The Web content or Text Mining can be distinguish from the two views namely Database View and Information Retrieval View. unstructured documents can be represented through the use suitcase of words. This representation ignores the order of the occurrence of words. The feature could be Boolean (i.e whether the word occurs or do not occur in the related article), or frequency based(ie how many times the particular word repeats). The features could be extracted by using some mining techniques such as cross entropy, mutual information or information gain [2]. Latent Semantic Indexing that translate the real document into lower dimensional space by observing the co-relational composition of the document cluster such that same documents that do not share terms would be placed together in the same category and stemming which reduces words to their morphological roots like “collection”, “collect”,“ collected”, “collecting” would be stemmed to their common root “collect” and only the latter word is used as the feature instead of the former four[6]. But on the other hand, Database Approach is mainly used to handle the unstructured data into the structured data by using related Data Mining techniques.

- **Multilevel Database:** This approach mainly focus on that the unstructured data of low level is stored in several web databases, like HTML documents. But the generalizations are made at the upper level which results into the organised structure [4].

- **Web Query Systems:** Database query language such as SQL is used by some web refer systems, eg.W3QL [4].

2.2. Web Structure Mining

The structured summary of web page and web site can be identified through web structure mining. The structured information is discoverable due to the availability of database techniques for web pages. Web Content Mining generally deals with the inner-document structure but, the Structure Mining focuses on the structure of the links of hyperlinks at the inter-document level mainly. The Web Structure Mining generally explains the web pages and produces the related information about the peculiar topic. Use of the Structure Mining reduces the two major issues of the web which occurs because of the abundant amount of data available [2]. The two major problems can be defined as following:

1. **Unrelated conclusion of search**: distortion occurs for corrected search as a result of search engines which allow for precision method.

2. **Availability of the abundant data:** Another problem is the indexing of large data quantity available on internet. The above reduction is a functional part of ascertaining the model beneath the web hyperlink structure given by the web structure mining.[3].

This mining extracts the unknown relationship between the web pages. This mining provides usage of link knowledge of one's own website endowing navigation as well as chunk of data into site maps. The relevant information can be promised with the use of keywords. Hyperlink command chain is decided to acquire related information within the sites as a relationship between competitor links and connection by means of third party co-link and search engines. Web Structure Mining also helps in establishing the similar structure of web pages by means of clustering technique. If there are huge web crawlers there will be more beneficial desired results to the related search [5].

If the web pages are directly linked with each another or web pages are neighbour we could find the relation among them. The relations may fall in category of ontology, they may have similar contents. Web Structure Mining also leads to generalize the sequence or networks of hyperlinks in the Websites in some particular domain. This leads to judgement of flow information in sites and this leads to easy and efficient query processing [1]. During 1997-1998, two most powerful hyperlink–based search algorithms Page Rank and Hits were introduced, which are HITS and Page Rank Algorithm and improvement of Hits by adding content information to the links structure and by using outlier filtering. These methods are mainly used to calculate the quality rank of each webpage.
Hyperlinks mainly act as useful for the following[1].

- Exploring the real pages link.
- Suggesting the pages with authority on the similar subjects the page containing the link.

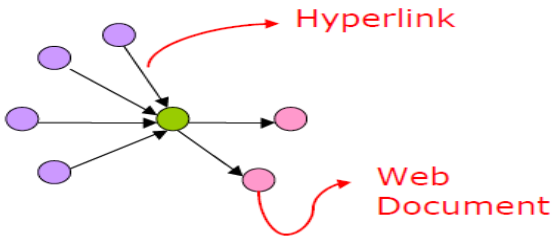

Fig 2. Web Graph Structure [6]

2.3. Web Usage Mining

The Web Usage mining is the third type of categorization. This permits us to gathering information for web pages. The related information is collected automatically into access logs. Mostly the organisations collect the daily logs who access the internet, for how much time, what sites he had visit via CGI scripts. By examining this organizations regulate promotional struggle, and customers life time[3]. Example online selling advertisement on the web. This mining also regulate the best path for the services[3]. Most existing tools provide the information about the user logs. By using such tools we can easily find out the details of user to determine that how many times one visit a particular site, name of the domain and the URLs of the user. But the traffic is handle from low to moderate side by these tools. More advanced systems are designed to discover and analysis of the patterns. The tools are differentiated into the following types namely:

1. **Pattern Discovery Tool:** There are some rising tools which are used to discover the patterns from techniques like Data Mining, Psychology and information theory ,Artificial intelligence to extract the knowledge from pool of information. Example WEBMINER System. It helps to discover association rule and sequential patterns from access logs automatically. [4]

2. **Pattern Analysis Tool:** After the finding of the patterns analysts wants the most suitable tool to envisage and interpret the patterns using the OLAP technique to discover the patterns [4]. eg. WebSIFT: It stands for Web Site Information Filter System**.** This Web Site Information Filter System act as a framework for web usage mining and use the structured information and content information to find the results. This is also fertilized research area [3]. Along creation of sever session, WebSIFT also execute content and structure pre-processing. The sequential pattern analysis, association rule discovery are performed on the session files to find the pattern analysis [3].

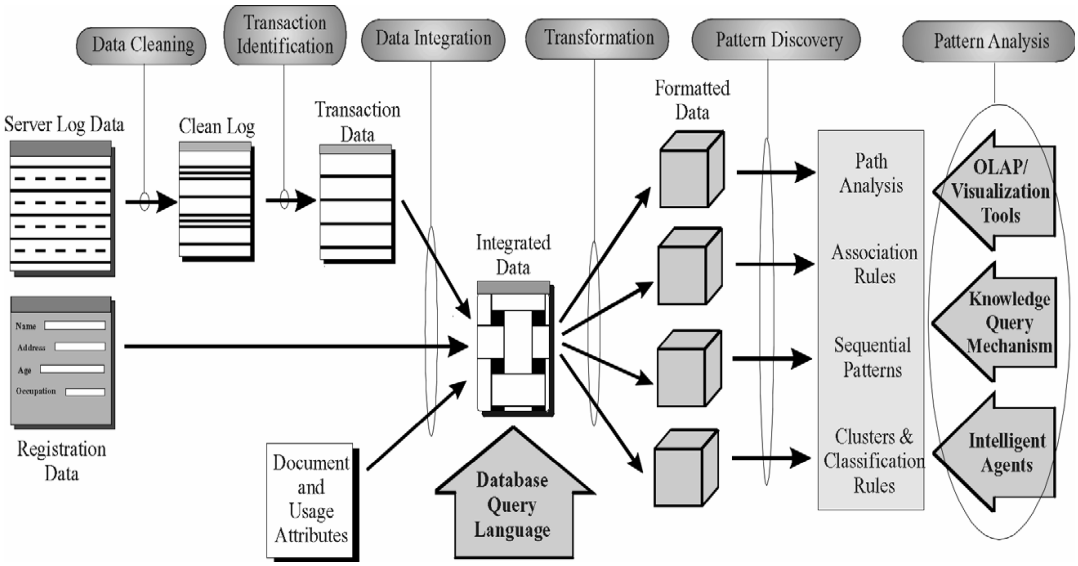

Fig 3. Web Usage Mining Process [4]

III. Analysis between the different categories of Web Data Mining

	Web Content Mining	**Web Structure Mining**	**Web Usage Mining**
Objective	Generally used to discover knowledge, collection of document like images and videos	Mainly structure of the link can be examined and also desirable documents can be determined	the behaviour of users can be determined during their interaction with World Wide Web
Technology Used	Machine learning and Automatic extraction	Information can be accessed through reference schema using database technololgy	Association, Clustering , classification
Data view point	Unstructured or semi-structured	Linked composition	associated
Application	Information extraction, segmenting web pages and detecting noise	Used in business to determine the link among various sites	Used in online auction, E-Banking, E-Commerce transaction, E-Commerce customer behaviour analysis

Table 1. Web Mining Categories [1]

IV. Conclusion

The Web Data Mining provides the way to categorize a required information. This mining helps to introduce the discovering of new tools to extract the valuable information from the sources. It gives the relevant results for the queries find out on the World Wide Web. The most related answers get find out through the use of techniques. In this work, Comparative analysis of the mining techniques has been done and it has been concluded that Web Content Mining has unstructured or semi- structure view of Data Mining and explains the automatic exploration of the information which could be accessible from the web . Whereas Structure Mining have linked structure at the inter- document level and helps to solve the problem of irrelevant results. The major use of the web structure mining is to find the hidden relationships between the Web pages. On the other hand, Web Usage Mining mainly includes interaction. Web Usage Mining provides the details of web access logs. The existing tools like WEBMINER helps to conclude the results of various web logs. There are many fields which could take the benefit of these mining categories. Web Usage Mining applications like E-banking, E-Commerce customer analysis and their problems require further research.

References

[1] Jaideep Srivastava, Robert Cooleyz , Mukund Deshpande, Pang-Ning Tan, "*Web Usage Mining:* Discovery and Applications of Usage Patterns from Web Data",*SIGKDD Explorations volume-1, Issue-2*, Jan 2000

[2] Raymond Kosala and Hendrik Blockeel," Web Mining Research: A Survey",*ACM SIGKDD*, July 2000

[3] Yan Wang," Web Mining and Knowledge Discovery of Usage Patterns*",CS 748T Project (Part I)*, February, 2000.

[4] R. Cooley, B. Mobasher, and J.Srivastava," Web Mining: Information and Pattern Discovery on the World Wide Web",*Ninth IEEE International Conference, IEEE*, Nov 1997.

[5] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava ,"Data Preparation for Mining World Wide Web Browsing Patterns",*Supported by NSF Grant*, Oct 1998.

[6] R. Kosala, H. Blockeel, "Web Mining Research: A Survey", *in SIGKDD Explorations 2(1), ACM*, July 2000.

[7] B. Masand, M. Spiliopoulou, J. Srivastava, O. Zaiane, ed. Proceedings of "*WebKDD2002 –Web Mining for Usage Patterns and User Profiles*", Edmonton, CA, 2002.

[8] R. Kohavi, "Mining E-Commerce Data: The Good, the Bad, the Ugly", *Invited Industrial presentation at the ACM SIGKDD Conference*, San Francisco, CA, 2001.

[9] M. Spiliopoulou, "Data Mining for the Web", Proceedings of the Symposium on Principles of Knowledge Discovery in Databases (PKDD), 1999.

Web References

[1] *https://sites.google.com/site/assignmentssolved/mca/semester6/mc0088/14*

*[*2] *http://www.web-datamining.net/*

[3] *https://www.google.co.in/*