ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Dec 2022

WEB MINING - SUSTAINABLE DEVELOPMENT IN DATA MINING

Dr. Trupti Sandeep Wani
Assistant Professor, SIES (NERUL) College of Arts, Science and Commerce
wani.trupti47@gmail.com

ABSTRACT

"Web Mining" - knowledge extractor from the huge repository i.e. World Wide Web. It uses the data mining techniques to extract the valuable, needful information from the internet. The various types of information are required to enhance the profit, to get popularity, to take wise decisions. All this type of data is gathered from web content, web structure and web usage data by using the web mining techniques. Because of providing large, very essential information, the web mining gets the popularity very fast among researchers and users. So in today's world, where required data is available and accessible to all easily, then Why not to use it for our benefit? In this paper web mining and its applications in real world for sustainable development are discussed.

Keywords: *Web mining, Knowledge Extractor, Data mining, Web structure, Web content, Web usage*

1. INTRODUCTION

Internet - repository of huge data. The information covers variety of aspects in digital format. Extracting the data from the repository and use of that data for taking decisions in the future is data mining. Web mining is the application of data mining. The data may be present in any form such as web documents, textual data, hyperlinks, audio, video, blogs, usage logs of web sites etc., the data mining techniques are used to extract knowledge from this. Different information retrieval and data mining techniques are used to extract knowledge. Web mining uses different traditional data mining methodologies and techniques to search information quickly and in easy way. It discovers the important data from World Wide Web. The relevant information is extracted from corpus by using machine learning and data mining techniques.

Process centric view and data centric view are the two approaches to define the web mining. First is the process centric means having the processes or sequence of tasks and another data centric where the web data is mainly consider for mining process. Actually web mining is the extracting the knowledgeable data by using many processes. So the main thing in web mining is extraction of web data to get knowledge. The web data means web content, web structure and web usage data.

2. LITERATURE REVIEW

Web data is in different format – structured or unstructured. Different researchers are classifying the web data in many ways. Cooley categorized the data as structure data, content data, usage data, profile data[4]. S.K.Madria et al. classified data in three types on the web, web log and web structure data[7]. Raymond also suggested three categories of web mining [8].The three categories of web mining web text mining, web usage mining and user modelling mining are suggested by Spiliopoulou[5]. The method of Latent Semantic Indexing is suggested by researchers of Cornell University to ascertain the topic words.[11].

Many new concepts are introduced by the researchers in web mining such as Hub and authority, page rank, Separation of human and non-human behaviour, browsing behaviour of user, understanding user by his profile, interest measure, sentiment analysis, pre-processing of data, visualization of WWW, online bibliometrics and so on.

3. Taxonomy of Web Mining

There are three different categories of web mining viz. web content mining, web usage mining and web structure mining. All these categories are depending upon the kind of data which is to be mined.

3.1 Web Content Mining

Extracting the data from web pages or documents i.e. the contents of web pages are known as web content mining. Web pages mainly contain the facts. Facts are in the form of text, images, audio and video or in structured forms such as tables or list. In web pages mining there are many issues such as extracting and discovering relevant information, finding association patterns, classification or clustering of data and so on. So there is a vast scope for the researchers to research in this area. As it is purely related with the data so the main focus of researchers are developing different techniques or bringing improvement in Information Retrieval (IR) and Natural Language Processing (NLP) techniques. Significant amount of work is already done in the area of image processing but the applications of those are inadequate in web content mining.

3.2Web Usage Mining

Searching interesting patterns from web usage data is known as web usage mining. This data is useful for finding and understanding the needs of web based applications. In web usage mining the origin or login identity of users is captured, also their browsing behaviour is taken into consideration. It is classified as - a) Web Server data, b) Application Server data and c) Application Level data. This classification is depending upon which data is used for consideration.

- Web Server Data – Users login data is mainly consider here. It includes IP address, access time and page reference.
- Application Server Data – The applications such as e-commerce application requires data such as web logic, designing, story server etc. Various kinds of business events are tracked and log them in application server data.
- Application Level Data – To generate the histories the logging information of new events are recorded and stored in application level data.

3.3Web Structure Mining

Discovering structural information from web pages is known as web structure mining. It consists of graphs having nodes and connecting edges. Nodes represent web pages and edges represent hyperlinks. There are two types of structures used in graph.

a) Hyperlinks – The hyperlinks which are present in the webpages are considered as structural unit. The hyperlinks are of two types - inter document hyperlink where link is in between two different documents and intra document hyperlink where link is within the document means same page at different location.

b) Document Structure – The web pages are formed by using HTML and XML tags. This information can be represented in the form of tree like structure and utilized for web mining.

4. Applications of Web Mining

In last few years the usage of World Wide Web increased drastically. The techniques increases in development of the web pages and use of the available information on net also increases. Few years back the small scale as well as large scale companies and industries are using RDBMS to handle the activities, transactions and generate the revenues but now they are using the web mining techniques to increase the revenues. It means there is a large development in web applications and it increases much faster. Not only in single one area but in all fields the web applications are used. They are using the web mining techniques and also bringing some modifications in the existing techniques and using it. Here in this section few successful applications are described which uses web mining to increase revenue or to take decisions.

- **Search Engines**

The main job of search engines is to provide relevant, correct information to user in small amount of time. The most famous search engine is Google. So here we will see how efficiently Google uses web mining techniques to bring up gradation in the existing system.

The most widely used popular search engine is Google. After getting users query it searches over 2 billion documents which are stored in its server. The documents are indexed and displays whenever relevant. How quickly and efficiently it searches the required information gives the popularity to the search engine. For the quality results Google uses structural information of the web graph. The first who introduces the importance of link structure from web mining in search engine is Google. The priority means importance of web pages are decided by using page rank algorithm by Google.

Many different facilities are provided by Google such as Google Toolbar which helps to search easier. Also additional information is provided by Google is that it highlights the query words in presented documents. It also captures the 'click stream information' if the full version of Google toolbar is installed. The captured information is used to enhance the facilities. The advanced search capabilities are used to find images and pages having required information in the data range. To increase the number of users Google also displays the related advertisement so that number of clicks increases. A leading national publication recognizes the Google as the most powerful and targeted B2B (Business-to –Business) advertising outlet.

'Google News' another popular facility provided by Google. It displays current, latest news from world. On regular basis it updates the news by using the algorithm from various sources. So there is no biasness in displaying the news. To read most relevant news, it incorporates news from various online news papers and arranges categorically.

- **B2C (Business-To-Customer) E-commerce**

To increase the profit in B2C, the idea of online store came and now they are generating a huge profit by analysing customer's behaviours by using web mining techniques. The best examples are Amazon, Mantra, IndiaMart , NyKaa and so on.

It is said that the CEO of Amazon Jeff Bezos observed that "In a traditional store, the main effort is in getting a customer to the store. Once a customer is in the store they are likely to make a purchase – since the cost of going to another store is high – and thus the marketing budget is in general much higher than the in-store customer experience budget (which keeps the customer in the store). In the case of an on-line store, getting in or out requires exactly one click, and thus the main focus must be on customer experience in the store." This fundamental idea encourages in starting the online business.

The web mining techniques such as click path analysis, association between visited pages, sentiment analysis etc. helps to understand customer's behaviour and to gain knowledge out of existing information.

- **Customer's Behaviour**

Customer's behaviour is very much important when the decisions have to take in the departmental store. The requirements are changing depending on weather, festivals, holidays, needs, like and dislike. Always the owner has to study this and has to take wise decision to increase the sale and revenue. Understanding customer's behaviour is now just one click away. As the trend of online shopping increases, it's easily possible to record the each activity of customer. It gives the detailed insight of user profile which helps to take wise decision. The sentiment analysis of web mining technique helps to do the analysis of customer behaviour.

- **Interestingness Measures**

As there is consumer behaviour recording and analysis is possible by using web mining, another very important application of web mining is measuring interest of users who are using internet, visiting sites and reading from websites. The readers are getting huge amount of information online so there is no need to buy books or to go to the library and search the required books. This facility is very much useful to researchers, as the researchers are getting vast amount of information of their interest. They can see and refer different publications or can themselves publish their research work. Researchers are getting in-depth knowledge of their interested area. Because of this it may possible that they can do direct interaction with other researchers and can innovate something good. It also helps the researchers to check that the material available on net is authentic or not. How many people visited the page, how much time they spent on that page, link structure and so on. By observing all these proper analysis is possible which was not possible earlier when so much online things and web mining is not available.

Like this there are so many applications of web mining. The complete procedure of web mining includes - searching of relevant data, pre-processing of data and visualization of results. All these things are possible because of different web mining techniques. After visualization it's very easy to take decisions.

CONCLUSION

Internet and its usage increases day by day. The data present on WWW is huge and growing drastically. To increase the business, wise decisions has to take and for that some useful information is required. Web mining provides the process and techniques to extract the knowledgeable data from WWW. Different types of information can be extracted by using different techniques from different web pages. So web mining helps in different areas to gather required information. Many web mining techniques are available to extract data, to analyse the data and also to take wise decisions. So web mining becomes the essential technique and helps in sustainable development in data mining process.

REFERENCES

1. Y. Yang, C. G. Chote(1994), An example-based mapping method for text categorization and retrieval. ACM Transaction on Information Systems (TOIS), 1994, 12(3): 252-277 32.
2. Yang: Expert Network(1994), Effective and efficient learning from human decisions in text categorization and retrieval. Procedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SIGIR'94), 1994: 13-22
3. Ester, H. P. Kriegel, J. Sander, X. Xu(1996), A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceeding of the 2nd Internatioal Conference on Knowledge Discovery and Data Mining, 1996: 226-231
4. R. Cooley: Web Usage Mining: Discovery and Application of Interesting Paterns from Web data. PhD thesis, Dept. of Computer Science, University of Minnesota. May, 2000
5. M. Spiliopoolou: Data mining for the Web. In Proceedings of Principles of Data Mining and Knowledge Discovery. Third European conference, 1999, 588-589
6. T. Mittchell: Machine Learning. McGraw: Hill, 1996
7. S. K. Madria, S. S. Rhowmich, W. K. Nig, F. P. Lim: Research issues in Web data mining. Proceedings of Data Warehousing and Knowledge Discovery, First International Conference. 1999: 303-312
8. Raymond Kosla, Hendrik Blockeel: Web mining research: a survey. ACM SIGKDD Explorations Newsletter, 2000, 2(1): 1-15
9. X. L. Li, J. M. Liiu, Z. Z. Shi(2000), The concept-reasoning network and its application in text classification. Journal of Computer Research and Development (in Chinese), 2000, 37(9): 1032-1038
10. Zhonghi Shi, Qing He, Ziyan Jia, Jiayou Li(2003), Intelligence Chinese Document Semantic Indexing System. International Journal of Information Technology and Decision Making, 2(3): 407-424
11. Radevr, Jing Hongyan, Budzikowska Malgorzata(2000), Centroid-based summarization of multiple documentsSentence extraction, utility-based evaluationand user studies. ANLP-NAACL 2000 Workshop, 2000: 21-29
12. A. Priadana and A. W. Murdiyanto(Jun. 2020), Analisis Waktu Terbaik untuk Menerbitkan Konten di Instagram untuk Menjangkau Audiens, J. Penelit. Pers dan Komun. Pembang., vol. 24, no. 1, pp. 59–70, doi: 10.46426/jp2kp.v24i1.118