ISSN 2348 - 2249
Volume-2, Issue-2, June-2014
International Journal for Multi Disciplinary Engineering and Business Management
(IJMDEBM)

Analysis of different Web Data Mining Techniques: A review

Pushpinder Kaur
Research Scholar, Department of Computer Engineering
Punjabi University, Patiala
India
pushpinder.kaur2@gmail.com

Himanshu Aggarwal
Department of Computer Engineering
Punjabi University, Patiala
India
himanshu.pup@gmail.com

Abstract **-** In this paper, we have reviewed different types of Web Data Mining techniques. The use of Internet or in other words World Wide Web (WWW)is increasing day by day. It comprises a huge amount of diverse data, thus it has become a big problem of extracting the required data easily and in a viable time. Thus we can describe Web Data Mining as a process of extracting required information from the World Wide Web efficiently, correctly, timely and also using minimum amount of resources [1]

Keywords - Data mining, World Wide Web, Extraction, Database.

I. INTRODUCTION

Due the abrupt increase in the users of World Wide Web and also the increase in the usage per user, the data available on internet has increased explosively. To search for the required data a user normally does it with the help of a few keywords. But with the increase in the data a user gets a large amount of data in front of him and it becomes very difficult for the user to select the appropriate data. To solve this problem Data Mining techniques are applied. Thus Data mining can be described as the extraction of relevant data from a huge amount of data. The main aim of the Data Mining is to extract the required information in such a form that a user can understand and also to present the information in a good way. As the World Wide Web contains huge amount of data and that too in different forms like Text, Images, Audio, Video etc., so to use this information properly and efficiently a good and efficient technique is required.[1][3]

II. II. LITERATURE REVIEW

Web Mining – An Integrated Approach:

In this paper, the authors have presented a trend discovery technique for dynamic web data mining. This technique canincrease the potential of traditional web data mining techniquesso that they should be able to analyse the frequently changing web sites that contain information like online news sites.According to the authors the data can be collected from 3 different locations:-

1. Server side collection - The browser behaviour of a particular web user can be collected from the log files of a site's server.

2. Client side collection –It makes use of a client side application, such as a remote agent to collect the data of the user navigation.

3. Proxy side collection –It is like the server side collection, proxy side collection, collects the data in a log file. It can be useful to characterize a group of users that use the same proxy server.

The technique proposed by authors consists of following things:-

1. Web log data preparation- Not every data log that is available is useful therefore we need to first prepare this data. Preparation of the data includes Data Cleaning, Efficient User Identification, Session Identification and Path Completion and Transaction Identification.

2. Web personalization- This process is used to identify the generation of a particular user model. It is based on the fact that different users rate different content for different reasons and under different contexts.

3. Detecting noise in the web pages- The noise should be detected and removed from a particular webpage so as make the data mining much faster and efficient.

4. Binding the web information- In order to make use of or to extract information from multiple sites there is a need to semantically integrate information from multiple sources.

This technique increases the potential of traditional web data mining techniques to analyse the frequently changing web sites.[4]

Extracting Data through Web mining:

In this paper the author has divided the Web mining process into the following categories:-

1. Web Usage Mining- It can be described as the process of extracting useful information from a server log file. In this we have to find what users are seeing on World Wide Web.

2. Web Content Mining- It can be described as the process of extracting relevant information from a web page.

3. Web Structure Mining- In this process we use the graph theory to analyse the nodeand connections between various structures of a website. In this process different web pages are denoted as a nodeand hyperlinks are denoted as an edge.

ISSN 2348 - 2249
Volume-2, Issue-2, June-2014
International Journal for Multi Disciplinary Engineering and Business Management
(IJMDEBM)

The author has also explained the pros and cons of web mining and has given a deep analysis of difference between web mining and data mining.[1]

A Brief Survey of Web Data Extraction Tools:

In this paper the authors have presented a deep survey of tools that are used for the generation of wrappers to extract data from sources. The wrapper development is commonly done by code writing in languages. The authors have presented the taxonomy to classify these tools according to the technique they use to generate wrappers. The taxonomy given by authors is as follows:-

1. HTML aware Tools- These tools make use of HTML parser that constructs a parsing tree following the document object model. HLML aware tools are XWRAP, Road Runner, W4F etc.
2. NLP based Tools- These tools aims at extracting data from free text. These take an input as a document and a filled template that indicates the data that is to be extracted. Examples of these types of tools are SRV, WHISK etc.
3. Wrapper Induction Tools- These tools take as input a set of pages where data of interest is labelled to serve as an example.These include WIEN, SoftMealy, STALKER etc.
4. Modeling-based Tools- These include tools that given a target structure for objects of interest, try to locate in web pages portions of data that implicitly conform to that structure. These include NoDoSE, DEByE etc.
5. Ontology based Tools-In these ontologies are previously constructed to describe the data of interest.

The authors have also analysed these tools qualitatively by examining how they support the most important features for the generation of wrappers and also the process they use while extracting data.[3]

Design and Implementation of A Web Mining Research Support System:

The authors have proposed a designed and planned implementation of a web mining research support system. This system is designed in such a way that it is able to identify, extract, filter and analyse data from web resources. The proposed system is composed of many stages which are Information Retrieval (IR), Information Extraction (IE), Generalization, and Analysis& Validation. The system can provide a general solution which researchers can follow to utilize web resources in their research. IR is used to identify the web sources with the help of predefined categories with automatic classification. IE uses a hybrid extraction in such a way so that it can select some portions from a web page and put those portions into the databases.

Generalization cleanses the data and uses the database techniques to analyse the collected data, Simulation and Validation builds models based on the extracted data and also validates their correctness. This system offers an integrated set of web mining tools that can help researchers to do online research[2]. Web usage mining result can be improved by analyzing web content. The system integrates web page clustering into log file association mining and cluster labels are used as web page content indicators. The Web page clustering was done using K-means algorithm. The clusters obtained from the web log file and integrated data file were manually summarized.

III. FUTURE SCOPE

There is a lot of future scope in Web Mining. It can help a lot in ecommerce due to the ease in personalizing marketing which can eventually provide a very higher volume of trades. It can also help in countering the Cyber Terrorism. Web Mining is also a powerful tool for Cloud Computing. Thus, if we get a good technique for web mining we can revolutionize the world. The work proposed in this project has lots of future scope because we will concentrate on the different techniques of web content mining. Web content mining has been proved very useful in the business world. The survey regarding these methods also shows the techniques used for extracting information from different types of data available in the internet and how this extracted data can be used for mining purposes. Users feel difficulty in finding desired information and deciding which information is relevant to them from general purpose search engines. Web content mining solves this problem and helps the users to fulfil their needs.

IV. CONCLUSION

The investigations have shown that by using different proposed techniques and algorithms we can bring a lot of improvement in Web Data Mining techniques. Different techniques showed improvement in different parameters. Thus by selecting appropriate techniques for different purposes we can really improve the quality of Web Data Mining.However, the actual tool which will be used for data extraction depends upon the requirement of the user.

REFERENCE

[1] Mrs. BhanuBhardwaj, “Extracting Data Through Webmining”, International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 3, May - 2012.
[2] Jin Xu, “Design and Implementation of A Web Mining Research Support System”.
[3] Alberto H. F. Laender, Berthier A. RibeiroNeto, Altigran S. da Silva and Juliana S. Teixeira,”A Brief Survey of Web Data Extraction Tools”
[4] N. Senthil Kumar P.M. Durai Raj Vincent, “Web Mining – An Integrated Approach”,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 3, March 2012
[5] B. Adelberg “NoDoSE-a tool for semi-automatically extracting structured and semi structured data from text documents”, p 283-294, 1998.
[6] R. Albert, H. Jeong, and A. L. Barabasi, “Diameter of the world wide web”, p 130-131, 1999.
[7] A. A. Barfourosh, H.R. MotaharyNezhad, M. L. Anderson, D. Perlis, “Information Retrieval on the World Wide Web and Active Logic: A Survey and Problem Definition”, 2002.
[8] Cooley, R.; Mobasher, B.; Srivastava, J.; “Web mining: information and pattern discovery on the World Wide Web”.

InProceedings of Ninth IEEE International Conference. pp. 558 –567, 3-8 Nov. 1997.
[9] B. Singh, H.K. Singh, "Web data Mining Research", in the Proceedings of 2010 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-10, Dec. 2010
[10] O. Etzioni, "The world wide Web: Quagmire or gold mine." Communications of the ACM, Vol. 39 No. 11, pp. 65-68, Nov. 1996.
[11] G. Salton and M. J. McGill, "Introduction to Modern Information Retrieval", McGraw-Hill, New York, 1983.
[12] C. Shahabi, A. M. Zarkesh, J. Adibi, and V. Shah, "Knowledge discovery from users web-page navigation" In Workshop on Research Issues in Data Engineering, Birmingham, England, 1997.
[13] S. Soderland, "Learning information extraction rules for semi-structured and free text".

