Research Issues in Web Mining

Pallavi Kamalakar Bhoir
ME Scholar
Department of Computer Engineering
Shree L.R. Tiwari College of Engineering, Mira Road, Thane-401107
University of Mumbai
pallu.bhoir@gmail.com

Prof. Neha Jain
Assistant Professor
Department of Computer Engineering
Shree L.R. Tiwari College of Engineering, Mira Road, Thane-401107
University of Mumbai
nehajain20683@gmail.com

Abstract **-** Web has more powerful platform for retrieval different information and discovering knowledge from web data. Web is a collection of files on one or more web servers. Web mining means extracting information from web database. The web data includes web link, web pages, web logs and object on the web. Web mining is used to the customer behaviour and understands the customer. Web mining using different techniques classification, clustering and association rules.it has some application such as Electronic commerce-learning, E-government, E-policies, Electronic business, security, crime investigation and digital library. Retrieving the required web page information from web side becomes a difficult task because web is made up of large unstructured data, which delivers the huge amount of information and increase the complexity from different web service providers. Web mining can be classified into main tree areas: web usage mining, web content mining, and web structure mining.in this paper, we have studied the basic concept of web mining, and issues.

Index Terms **-** *Web mining, classification, Application, Tools, Algorithms, Research issues.*

*****

I. INTRODUCTION

The Web mining is the process of using data mining techniques and algorithms to extract information from web side. Actually web mining is the application of data mining technique which is two types of data that is an unstructured or semi-structured data and it is automatically extract useful information or knowledge from web [1]. In web mining different application are website design, web search engines, information retrieval, network management, e-commerce, artificial intelligence and business. this application includes the temporal issues for the users.

Web mining can be classified into main tree areas: web usage mining, web content mining, and web structure mining. Each classification is having its own algorithms and tools. Web content mining is nothing but text mining; it is generally the second step in web data mining. Web content mining is the scanning and mining of text, hyperlink, pictures and graphs of a web page. Web structure mining it is one of the three categories of web mining for data.in web structure mining is a tool used to identify the relationship between web page. Web usage mining is also called as web log mining. Actually web usage mining data captures the identity or origin of web users and browsing behaviour at a web side [2].

Web mining process consists of four important steps: Resource finding, Data selection and pre-processing, Generalization and Analysis. In the resource finding process which is used to extract the data two form either online or offline text resources. In the Data selection and pre-processing is the process specific information from retrieved web and automatically selected and pre- processed. In the generalization used data mining and machine learning techniques to discover general patterns from the individual web sites and multiple web sides. During Analysis step using validation and interpretation of the patterns are done. Web mining can be classified into main tree areas: web usage mining, web content mining, and web structure mining [3].

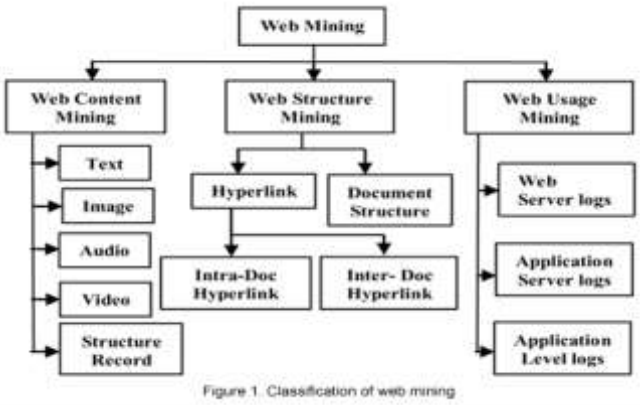


Figure 1. Classification of web mining

II. *RESEARCH ISSUSE IN WEB MINING*

The web mining is data sets can be very large. It can't handle ton signal server. In the web usage mining data set is very large so it is difficult to organize hardware and software [4]

A. Major issues in web mining

- Web data sets Can be very large, it takes ten to hundreds of terabytes to store on the database
- It cannot mine on a single server so it needs large number of server.

- Difficult in finding relevant information.
- Mining sequence and time service data.
- Automated data cleaning
- Limited query interface to individual users.
- Extracting new knowledge from the web.
- Scaling up for high dimensional data.
- Over fitting and under fitting of data.

II. *WEB CONTENT MINING*

The web content mining also called as the text mining. Web content mining is the scanning and mining of text, hyperlink, pictures and graphs of a web page.it is generally the second step in the web mining.

The web content mining process of retrieving the data from web into more structured forms and indexing the information or finding valuable information from web document or web content.in the web content mining different document such as text, html, multimedia documents that is audio, image video or sound the search result it may be a structure documents or unstructured document [5].

Web content mining used different algorithm and tools such as cluster hierarchy construction algorithm (CHCA), Genetic algorithm and web info Extractor, Mozenda are content mining tools. It has two approaches Agent based and Database Approach.

A. Agent Based Approach

The Agent based approach main focuses on searching the relevant information from the World Wide Web. Threes are three agents [6].

1. *Intelligent search agent*: the intelligent search agent automatically searches for information along with a particular query.
2. *Information filtering*: the information filtering also called as the categorizing agent it is filters the data.

B. Database Approach

The database approach consists of database which contains attributes, tables and schema with defined domains.

III. *RESEARCH ISSUSE IN WEB CONTENT MINING*

- Data/information Extraction concentrates on extraction of structured data from web pages such as products and search results.
- Web information integration and schema matching.

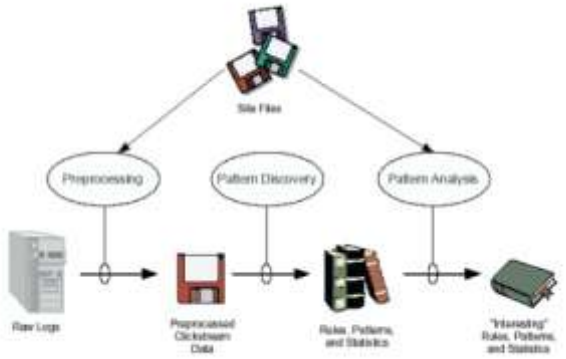
IV. *WEB STRUCTURE MINING*

The web structure mining is the data interconnected to the structure of a particular website.it is graph which is contain the web pages and web documents as nodes and hyperlinks as edges are connecting between two relates pages Web structure mining is a tool used to identify the relationship between web page. The web structure is to extract some interesting web graph patterns like co-citation, social, complete bipartite graphs.it is performed two level either intra-page or inter-page level. A hyperlink connect to a different part of the same page is called intra-page level or hyperlink. Web structure mining is used in search engines such as Google, Yahoo [7].

V. *RESEARCH ISSUSE IN WEB STRUCTURE MINING*

- Web structure mining has huge amount of data.
- Reducing irrelevant search result.
- Indexing information on the web

VI. *WEB USAGE MINING*



Web usage mining is also called as web log mining. Actually web usage mining data captures the identity or origin of web users and browsing behaviour at a web side. The web log is located in three different locations such as server log, web proxy server and client browser and it can be contain only plain text file (.txt).in the web usage mining large amount of irrelevant information are present in the web log file because it can be contains noisy information, incomplete information and unnecessary information[8].

A. WEB SERVER LOGS

1. *Error log:* This error occur when user click on a particular link on the browser but browser does not display the web page or website then the user receives 404 errors that are means page not found.

2. *Agent log:* The Agent log it is a standard log file and it also used to record of online user's behaviours, version, operating system.

3. *Access log:* The Access log is used to capture the information about the users and it has also many numbers of attributes.

4. *Referrer log:* Referrer log is used to store the information of the URLs of web pages on other sites

that link to web pages .that is if a user gets to one of the servers pages by clicking on a link from another site, the URL of that site will appear in this log[9].

B. PROCESS OF WEB USAGE MINING

The web usage mining process is generally divided into three tasks [10]:

1. *Data pre-processing:* Data pre-processing means cleans log files of server and also removing log entries such as user identification by ip address, authentication data, client information formatting, error repeated request for the same URL and the same host.
2. *Pattern discovery:* The pattern discovery also called as the knowledge discovery this knowledge collected can be used to take decision on various factors such as the Excellent,Medium,weak users and Excellent,Medium,weak web pages it is based on the hit count on web side.
3. *Pattern analysis:* The pattern analysis is main task in the web usage mining. it is also called as the last process of the web usage mining. There are so many different techniques are used for pattern analysis like visualization technique, OLAP techniques data and knowledge querying.

VII. *RESEARCH ISSUSE IN WEB USAGE MINING*

Web usage mining has several issues because number of data mining techniques.

- CGI data
- Catching
- Dynamic pages
- Session identification
- Transaction identification

VIII.*CONCLUSION AND FUTURE WORK*

In this paper we discussed about the different research issues and challenges in the web mining and also explain basic concept of web mining, web content mining, web structure mining, web usage mining, tools, algorithms and types.in this paper several open issues and drawbacks are exists in the current techniques are also discussed. This paper also helpful for researchers those are doing research in the web mining domain.

REFERENCES

[1] http://searchcrm.techtarget.com/definition/Web-mining
[2] http://www.ijcsmc.com/docs/papers/May2014/V3I5201461.pdf
[3] http://www.ijarcce.com/upload/2015/november-15/IJARCCE%2088.pdf
[4] http://www.ijcst.org/Volume3/Issue7/p15_3_7.pdf
[5] http://ijirts.org/volume2issue3/IJIRTSV2I3050.pdf
[6] http://www.ptc.org/ptc14/images/papers/upload/Paper_TS6YS1_Farhaan%20Mirza.pdf
[7] http://userpages.umbc.edu/~juacubi1/733/project/papers/Web%20Structure%20Mining.%20An%20Introduction.pdf
[8] http://yildiz.edu.tr/~aktas/courses/CE-0114890/g10-p3.pdf
[9] http://bipublication.com/files/IJCMS-V4I1-2013-01.pdf
[10] http://yildiz.edu.tr/~aktas/courses/CE-0114890/g10-p3.pdf