Web Content Mining and its Tools

C. B. Pavithra
Department of Computer Applications
KG College of Arts and Science

Abstract:- This paper presents the small print of searching or extracting information from the online . It also discusses main tasks involved in web mining. It mainly focuses on the kinds of web page mining like Unstructured, Structured and Semi-structured types. Finally, some tools that are used for mining is additionally focused during this paper.

Keywords:- *Web Mining, Web Content Mining, Structured Data extraction, Unstructured Data Extraction.*

I. INTRODUCTION

Data is that the prime source of processing in Computer. it will be in many forms and in many Places round the world. data are wasted if it's not processed and there are some ways to access the info for processing. Each way has its own hurdles in reaching users' knowledge. People get the info within the earlier days through the Newspaper and from Different Medias. Now- a-days the modern gadget Computer is used to quench the thirst of people Knowledge. WWW is the only popular medium to disseminate information to people. Many barriers within the work of retrieving pave the thanks to do research in searching the Web with the term 'WEB MINING'. Web pages are viewed using a browser. Deep web is accessing databases through queries.

Web mining is divided into four tasks:

- *Resource Finding:*
  Retrieving the document related to our search.

- *Information selection and Pre-processing:*
  Selecting an appropriate document from the displayed documents.

- *Generalization:*
  Discovering patterns from individual websites

- *Analysis:*
  Checking the validity of the mined patterns.

II. CLASSIFICATION OF DATA MINING

Data mining is broadly classified as Web content mining(WCM), Web Structure Mining (WSM) and Web usage Mining (WUM). WCM is identifying data for the user from the text, image, audio or video data. WSM is identifying node and connection to find the structure of a web site. WUM is identifying user access patterns from user logs.

III. CLASSIFICATIONOF WCM

Selected according to that, Data in web may be in any form such as table as structured data, free-text as unstructured data and HTML documents as types IR (Agent-based) view and DB (Database) view. With the DB view the data is combined and organized in such a way that sophisticated queries can be used for searching data in a database. In IR view helps to retrieve information easily from a drawing source of web data. The main object used in Web Content Mining is "Text documents".

The two main approaches in WCM are

- Structured data extraction and
- Semi-structured
- Unstructured data extraction.

- *Structured data extraction:*
  A large amount of information on the Web is contained in regularly structured data objects. Such Web data records are important because they often display the essential information on their host pages, e.g., lists of products and services. Two approaches used are: Wrapper Induction and Automatic data extraction. Many techniques have been used to retrieve appropriate information from the page. First, Classification of websites on top of web page demands new algorithms developed. Secondly, crawling of web page gives a large number of relevant data, Clustering is a method used to group the set of related information for better understanding. Crawlers used to traverse through structured data divides into internal which traverses through internal pages and external web crawler which traverses through unknown website.

- *Semi-Structured data:*
  The techniques used are OEM (Object Extraction Model) in which a group is formed which contains relevant information and stored in OEM, top down extraction in which complex objects are converted to atomic objects and web extraction language in which web data converted to structured data in the form of tables.

- *Unstructured Data extraction*
  Web content data is unstructured data and the research around it is knowledge discovery in texts (KDT). Some of the techniques are,

- *Information Extraction:*
  Keyword and phrases are identified and searched inside the text

- *Information Visualization:*

It utilizes feature extraction and key term indexing which builds image in large images.

- *Topic Tracking:*

In this method, user interest is checked from other profiles of user and topics

- *Summarization*:

Length of the document is reduced which helps the user to decide whether they need to semi structured data. Web content mining is classified as to read the document,

- *Categorization*:

Main themes are identified and number of words in that document is counted used for ranking that page.

- *Clustering*:

Similar documents are grouped which helps the user to select the topic of interest.

IV. WEB CONTENT MINING TOOLS

WCM tools help the user in downloading the information for users in an easier way.

Some of the tools are:-

➢ *Web content Extractor:*

This tool is used by book readers to extract details about books, businessman extract and collect price and real estate details, Journalists information, online information, Job seeking details.

➢ *Web Info extractor:*

It extracts unstructured data as well as tabular data to a file, Monitors web pages and retrieve new content from any kinds of file types.

➢ *Automation:*

Automates complex tasks, web recorder, automate scripts, powerful task scheduling and auto-run scheduled tasks.

➢ *Screen Scrapper:*

This tool allows us to Mine data on products and download them to a spreadsheet.

➢ *Mozenda:*

Agents are set up to extract data in a regular fashion and circulate it to several destinations.

➢ *SAS Enterprise Miner:*

User friendly GUI to the SEMMA (Sample, Explore, Modify, Model, Assess) process

➢ *QL2 Software:*

Data extraction using SQL query like language

➢ *Wizsoft Software:*

Software developed based on mathematical algorithms used in business sectors.

➢ *Website Parser:*

Helps to gather information from any website quickly, which helps for the business owner or retail sites.

V. CONCLUSION

The web is a largest data repository in the world. As the information is dynamic everyday new contents needs to be added on the web page. The way of searching information also differs in respect to people knowledge, the content of the page and necessity of them. New algorithms and techniques should be developed to cope with the growth of data repository. Data extraction should not completely depend on algorithms or techniques. Many tools have been developed for extracting information from the data warehouse which also contains some pros and cons that yet to be considered for further research.

REFERENCES

[1]. Arvind Kumar Sharma1, P.C. Gupta, “Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 1, Issue 8, October 2012.

[2]. V. Bharanipriya1 & V. Kamakshi Prasad2, ”WEB CONTENT MINING TOOLS: A COMPARATIVE STUDY”, International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4.

[3]. Mrs. Bhanu Bhardwaj, “Extracting Data through Webmining,” International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 3, May -2012 ISSN-0181.

[4]. Aidan Finn, Nicholas Kushmerick and Barry Smyth, “Fact or fiction: Content classification for digital libraries”, In Proceedings of Joint DELOSNSF Workshop on Personalization and Recommender Systems in Digital Libraries (Dublin),No. 01/W03 18 – 20, June 2001.

[5]. A. A. Barfourosh, H.R. Motahary Nezhad, M. L. Anderson, D. Perlis, “Information Retrieval on the World Wide Web and Active Logic: A Survey and Problem Definition”, 2002.