CURRENT RESEARCH ISSUES IN WEB DATA MINING : A SURVEY

Rajalekshmy K.D [1], *Dr. S. Sheeja* [2]

ABSTRACT

World Wide Web is the most widely known and biggest source of information for data mining research. So, it becomes a challenging task to extract useful and novel information and knowledge from this huge, dynamic structurally complex and ever-growing World Wide Web. Web data mining is broadly classified into three types: Web Content Mining, Web Structure Mining and Web Usage Mining. This survey outlines the current research issues in Web Data Mining.

Keywords : Web Data Mining, Web Content Mining, Web Structure Mining, Web Usage Mining, World Wide Web, research issues

I. INTRODUCTION

In Today's world, people rely on World Wide Web to send and receive information, to distribute their knowledge, for conducting online businesses, to express their opinions and views, to discuss with the people all around the world, for entertainment etc. Thus World Wide Web has become the largest source of information for mining tasks. The information present on the web is useful depending upon the choices and preferences of people searching for information. For example, for a person who is engaged in steel industry business, the information related to other domains may seem less worthy. Even though almost every kind of information is present on the web, it becomes a challenging task to extract the information based on user's preferences and interests. Web Data mining is the application of data mining which uses advanced data mining algorithms and techniques to extract interesting and potentially useful patterns from the web data. [2]

II. WEB MINING

Traditionally most of the data mining techniques are used for homogeneous data, but the data present on the web are heterogeneous such as text, image, audio, video, hyperlink, log files, etc. Therefore, traditional data mining techniques are not adequate for the data on the web. So, researchers combine more advanced data mining techniques with web resulting in a technique called wed data miming. Web data mining is a data mining application which is used to discover new and useful patterns from web activities and web documents.

Basic process of web mining is almost similar to the data mining process shown in figure.1 [2]

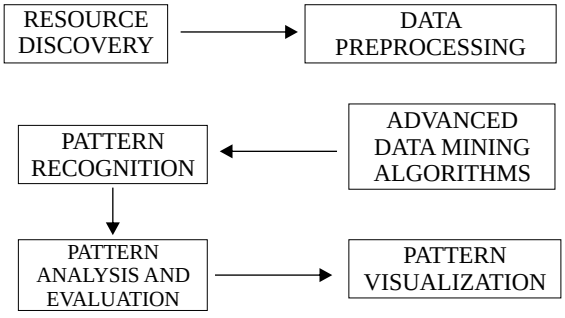

Figure 1. Basic process of web mining

[1]Research Scholar, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India

[2]Associate Professor, Department of CS, CA & IT, Karpagam Academy of Higher Education, Coimbatore, India

1. Resource Discovery: Discover unfamiliar documents, websites and web services on the web.
2. Data Pre-processing: Automatically extracting information relevant to our application by applying different pre-processing methods like feature reduction, feature subset selection, feature creation, sampling, transformation etc. from newly discovered web resources.
3. Advanced data mining algorithms: By applying advanced algorithms they uncover general patterns at individual websites and across multiple sites.
4. Pattern Recognition: Finds out the interesting and useful patterns from the general patterns.
5. Pattern analysis and Evaluation: Evaluates the patterns for accuracy.
6. Pattern Visualization: Represents the valid pattern in an easily understandable fashion.

Web data mining can be classified into three categories as shown in figure 2. Web Content Mining, Web Structure Mining and Web Usage Mining.

III. WEB CONTENT MINING

It is the process of extracting and integrating useful information and knowledge from web page contents. Web page consists of data in the form of text and multimedia, which are heterogeneous in nature. Most of the data present on the web are not in structured format. Data usually appear in semi-structured and unstructured format. There are several searching and indexing tools available on the Internet to cluster all similar and related pages and to distinguish home page from other pages based on a user query. However, they do not generally provide information in the structural format. This has prompted many researchers to develop new and more intelligent algorithms and techniques for information retrieval and organizing and interpreting semi structured and unstructured data.

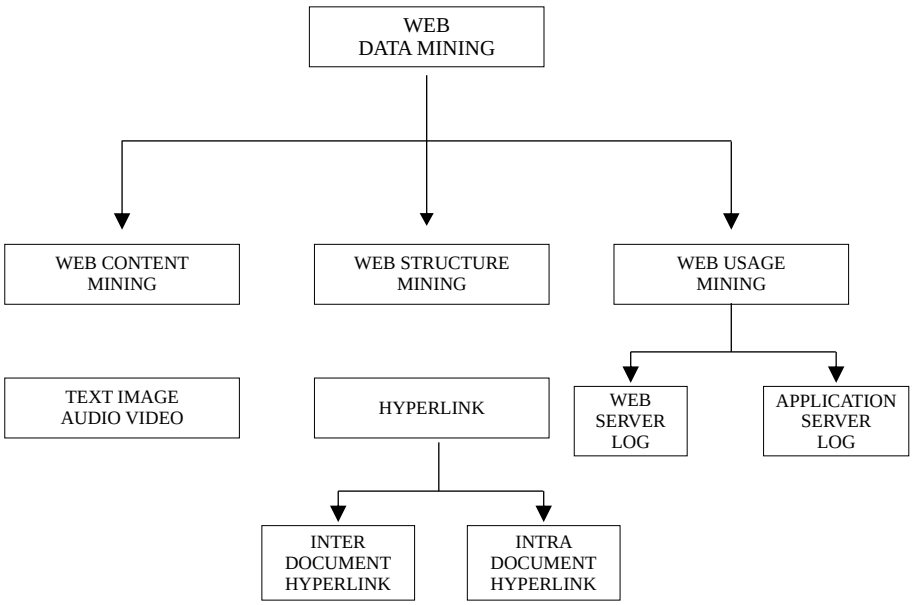

Figure 2. Taxonomy of Web Data Mining

The web content mining follows two approaches i.e. Agent Based Approach and Database approach. [1]

A. Agent based approach

This approach is used to discover relevant and significant information from World Wide Web automatically. It includes three types of agents. They are

Intelligent Search Agents: It searches for information automatically besides a particular query.

Information filtering/categorizing Agents: filters the data present on the web.

Personalizes Web agents: It discovers those documents which are anyhow related to the user profiles.

B. Database approach

It consists of databases which contain attributes, tables and schema with defined domains. The major effort is to organize semi-structured data on web into structured collection of resources and by using standard query or techniques to get effective results. Web Content Mining also employs other approaches like text mining, multimedia mining, structure mining etc.

IV WEB STRUCTURE MINING

It is the process of discovering useful patterns from the link structure of the web. Web information retrieval tools make use of only the text available on web pages, but ignore valuable information contained in web links. Web structure Mining aims to generate structural summary about websites and web pages. It uses graph topology to describe the website structure where websites represent nodes, and the links represent the arc connecting two interrelated web pages. Based on the type of web structural data, web structure mining can be divided into two types: [3]

" Hyperlink analysis: Extracting patterns from hyperlinks in the web. A hyperlink is a structural component that connects a web page to different web pages.

" Mining the document structure: Analysis of the tree-like structure of page structures to describe XML or HTML tag usage.

Web structure mining helps users to retrieve documents by analyzing the link structure of the web. This result in a newly emerging research area called link mining, which uses Link analysis Ranking algorithm to

1. Start the algorithm with a collection of web pages to be ranked.
2. Extract the hyperlinks between pages.
3. Construct the underlying graph.
4. Have this Graph given as input.
5. Assign a weight for each page where weight captures authoritativeness of the page and
6. Use authoritativeness to rank each page

Algorithms that best discover authoritative nodes in the graph must be devised where "authorities" are highly ranked pages for a given topic. [3]

V WEB USAGE MINING

Web Usage mining is the process of discovering interesting and meaningful usage patterns from data on web server logs in order to better serve the requirements of web-based applications. A user's identity and his browsing behavior can be captured from the usage data. Web usage mining consists of three steps [6]

1. Pre-processing: Removes noisy data and reduce the size of the data.
2. Pattern discovery: Cleaned log file is used to discover web usage pattern.

3. Pattern Analysis: Analyzing patterns in order to extract more useful information.

Web usage mining has several applications in e-business, including personalization, traffic analysis and targeted advertising. Main areas of research in this domain are web log data pre-processing and identification of useful patterns from this preprocessed data using mining techniques. Most data used for mining is collected from web servers, clients, proxy servers or server databases all of which generate noisy data. So data cleaning is necessary for click streams data. These data allow reconstruction of user navigational patterns. Personalization is one of the most widely researched areas in web usage mining.

VI ISSUES IN WEB DATA MINING

*** Large volume of data** : web data sets can be very large. It takes ten to hundred terabytes to store them on the database

*** High velocity of the data** : Data on the web are dynamic. It is very difficult to keep up with the speed with which they are changing.

*** Heterogeneous data** : Different varieties of data coexist on the web.

*** Hardware and software Management** : Proper organization of hardware and software to mine multi terabyte data sets is required which is not easy to manage. [1]

*** Data cleaning** : Automated data cleaning is required on large scale to find out useful information from data. [1]

*** Relevant information** : Difficult to find relevant information from large data sets. [1]

*** Fraudulent message detection** : It is difficult to find the authorship of the false and misleading messages that are being spread on the social media websites.

*** Limitations of the search engines** : Search engines show several limitations while performing searching for the user queries.

VII CONCLUSION

In this paper some basic concepts of web data mining have been introduced. This paper also discussed the categories of web data mining and some of the current research issues faced by web data mining researchers.

REFERENCES

[1] Kavitha, Priyanka Mahani Dr. Neelam Ruhil, "Web Data Mining A Perspective of Research issues and challenges," Internatinal Conference on computing for Sustainable Global Development(INDIACom) 2016.

[2] Brijendra Singh, Hemanth Kumar Singh," Web Data Mining research: A Survey"IEEE2010.

[3] R. Munilatha, K.Venkataramana, "A study on issues and techniquesof web mining" IJCSMC, vol.3, Issue 5,May 2014ISSN 2320-088X.

[4] Amarjeet singh Yumnam, Y. Chaithanya Sreeram, Shaik Abdul Naeem, " Overview: Web log miming, Privacy issues and application of web log mining", . Internatinal Conference on computing for Sustainable Global Development(INDIACom) 2014.

[5] Pranam Kolariand Anupam Joshi," Web mining : Research and practice",Computing in science and engineering 2004.

[6] Sunena, kamaljit Kaur,"Web usage mining - current trends and future challenges", International conference on Electrical ,

Electronics, and optimization Techniques (ICEEOT)-2016.

[7] Gustavo Rossi," A Survey of Web Research in Argentina",IEEE computer society.

[8] Jai Prakash Verma, Dr. Atul Patel, "Web Warehouse:Issues and Challenges for Web Data Mining", International Journal of Advanced Research in Computer Science Volume 8, No.5 May-June 2017.

[9] Bing Liu,"Web Data Mining-Exploring Hyperlinks, Contents and Usage Data", Second Edition Springer 2011.

[10] Jiawei Han and Micheline Kamber, " Data Mining Concepts & Techniques", Second Edition, Elsevier.