WEB MINING TECHNIQUES IN THE AREA OF THE WEB PRESONALIZATION

[1]NIDHI RAJ, [2]N.K.SINGH

[1]B.Tech (VI-Sem)., Department of Computer Science & Engineering, S.B.M. Jain College of Engineering, Jain University, Bangalore-560004, India.
[2]University Department of Physics, V.K.S.University, Ara, Bihar. India.
E-mail: [1]nidhidsi2013@gmail.com, [2]singh_nk_phy27@yahoo.com

Abstract - "Web mining refers to the overall process of discovering potentially useful and previously unknown information or knowledge from the Web data." With the large amount of information available online, the Web is a fertile area for data mining and knowledge discovery. Data mining has become so many key features for detecting fraud, assessing risk, and product retailing, In Web mining, data can be collected at the server-side, client-side, proxy servers, or obtained from an organization's consolidated web data. web mining can be categorized into three areas Web Content Mining, Web Structure Mining and Web Usage Mining. Web mining is the application of data mining techniques to extract knowledge from web data, where at least one of structure (hyperlink) or usage (web log) data is used in the mining process (with or without other types of web data)

Index terms - Web usage mining, Data mining, Web mining techniques, Personalization mechanism.

I. INTRODUCTION

Today the evolution of the World Wide Web has brought us enormous and ever growing amounts of data and information. It influences almost all aspects of people's lives. In addition, with the abundant data provided by the web, it has become an important resource for research. Furthermore, the low cost of web data makes it more attractive to researchers. Researchers can retrieve web data by browsing and keyword searching [1]. However, there are several limitations to these techniques. It is hard for researchers to retrieve data by browsing because there are many following links contained in a web page. Keyword searching will return a large amount of irrelevant data. On the other hand, traditional data extraction and mining techniques cannot be applied directly to the web due to its semi-structured or even unstructured nature. Web pages are Hypertext documents, which contain both text and hyperlinks to other documents. Furthermore, other data sources also exist, such as mailing lists, newsgroups, forums, etc. Thus, design and implementation of a web mining research support system has become a challenge for people with interest in utilizing information from the web for their research. A web mining research support system should be able to identify web sources according to research needs, including identifying availability, relevance and importance of web sites; it should be able to select data to be extracted, because a web site can be viewed as the largest database available and presents a challenging task for effective design and access.

II. WEB MINING

Web mining is a technique (Table 1) to discover and analyze the useful information from the Web data. According to Etzion [2], web mining can be divided into four subtasks:

i) Information Retrieval/Resource Discovery (IR): Find all relevant documents on the web. The goal of IR is to automatically find all relevant documents, while at the same time filter out the non-relevant ones. Search engines are a major tool people use to find web information. Search engines use key words as the index to perform query. Users have more control in searching web content. Automated programs such as crawlers and robots are used to search the web. Such programs traverses the web to recursively retrieve all relevant documents. A search engine consists of three components: a crawler which visits web sites, indexing which is updated when a crawler finds a site, and a ranking algorithm which records those relevant web sites. However, current search engines have a major problem - low precision, which is manifested often by the irrelevance of searched results.

ii) Information Extraction (IE): automatically extract specific fragments of a document from web resources retrieved from the IR step. Building a uniform IE system is difficult because the web content is dynamic and diverse. Most IE systems use the \wrapper"[3] technique to extract a specific information for a particular site. Machine learning techniques are also used to learn the extraction rules.

iii) Generalization: discover information patterns at retrieved web sites. The purpose of this task is to study users' behavior and interest. Data mining techniques such as clustering and association rules are utilized here. Several problems exist during this task. Because web data are heterogeneous, imprecise and vague, it is difficult to apply conventional

clustering and association rule techniques directly on the raw web data.

(iv) Analysis / Validation: analyze, interpret and validate the potential information from the information patterns. The objective of this task is to discover knowledge from the information provided by former tasks. Based on web data, we can build models to simulate and validate web information.

III. WEB MINING TASKS

Table 1

	Web Mining			
	Web Content Mining		Web Structure Mining	Web Usage Mining
	IR view	DB view		
View of data	Unstructured , Semi-structured	Semi-structured Web site as DB	Links structure	Interactivity
Main data	Text Document , Hypertext document	Hypertext document	Links structure	Server Log, browser Log
Representation	Bag of words, n-grams, Terms, Phrases, Concepts or ontology, relational	Edge- labelled graph (OEM), Relational	Graph	Relational table, Graph
Method	TFIDF and variants, Machine learning, Statistical	Proprietary algorithms, ILP, association rules	Proprietary algorithms	Machine learning, Statistical, association rules
Application Categories	Categorization, clustering, Finding extraction rules, finding patterns in text, user modelling	Finding frequent sub structures, web site schema discovery	Categorization, Clustering	Site construction, adaption, and management, Marketing, user modelling

Web mining tasks can be divided into several classes. Table 1 shows different categories of Web mining tasks. In web content mining as per IR view task based on unstructured and semi structured data where as the main data is text document and hypertext document. The method used in this process is machine learning, variants and statistical. The application based on categorization, clustering, finding extraction rules, finding patterns in text and various user modeling. In web structure mining task based on link structure data where as the main data is also linked structure. The method used in this process is proprietary algorithms and application based on categorization and clustering. In web uses mining data collected from user interaction whereas main data collected from server log and browser log. The method used in this process based on machine learning, statistical and association rules.

IV. WEB CONTENT MINING

Web Content Mining [8] is the process to describe the discovery of useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, metadata, hyperlinks or structured records such as lists and tables [**5**]. Research in web content mining encompasses resource discovery from the web, document categorization and clustering, and information extraction from web pages [6].

V. WEB STRUCTURE MINING

Web structure mining tries to discover the model underlying the link structures of the Web. This model can be used to categorize web pages and is useful to generate information such as the similarity and relationship among Web sites. Web structure mining studies the web's hyperlink structure. It usually involves analysis of the in-links and out links of a web page, and it has been used for search engine result ranking. [6]. Web Structure Mining can be regarded as the process of discovering structure information from the Web. This type of mining can be performed either at the (intra-page) document level or at the (inter-page) hyperlink level [5]. Web structure mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web [**7**].

VI. WEB USAGE MINING

Web usage mining (also referred to as click-stream analysis) [10] is the process of applying data mining techniques to the discovery of usage patterns from Web data, and is targeted towards applications [9]. It

tries to make sense of the data generated by the Web surfer's sessions or behaviors. While the web content and structure mining use the real or primary data on the web, web usage mining mines the secondary data derived from the interactions of the users during Web sessions. Web usage data includes the data from web server access logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries, mouse clicks, and any other data as the result of interaction with the Web. Area of Web Usage Mining: Personalization [4], System Improvement, Site Modification, Business Intelligent, Usage Characterization. A personalization mechanism is based on explicit preference declarations by the user and on an iterative process of monitoring the user navigation, collecting its requests of ontological objects and storing them in its profile in order to deliver personalized content.

CONCLUSION

In this paper, first we have mainly focused on the web mining task- IR will identify web sources by predefined categories with automatic classification. IE will use a hybrid extraction way to select portions from a web page and put data into databases. Generalization will clean data and use database techniques to analyze collected data. Simulation and Validation will build models based on those data and validate their correctness. After that, we have introduced the web mining types - Web content mining, web structure mining and web usage mining. After that, we have introduced the web mining techniques in the area of the Web personalization.

REFERENCES

[1] A. Laender, B. Ribeiro-Neto, A. Silva and J. Teixeira:"A brief survey of web data extraction tools. In SIGMOD Record", 2002, 31.
[2] O. Etzioni, The world-wide web: Quagmire or gold mine? Communications of the ACM, 39(11):65-68, 1996.
[3] L. Eikvil, Information extraction from World Wide Web - a survey. Technical Report 945, Norweigan Computing Center, 1999.
[4] D. Antoniou, M. Paschou, E. Sourla and A. Tsakalidis," A Semantic Web Personalizing Technique The case of bursts in web visits," presented at IEEE Fourth International Conference on Semantic Computing, 2010.
[5] A. J. Ratna kumar, "An Implementation of Web Personalization Using Web Mining Techniques, "Journal of Theoretical and applied information technology, 2005.
[6] W. Bin and L. Zhijing, "Web Mining Research," in Proceedings of the fifth International conference on Intelligence and Multimedia Applications (ICCIMA'03),2003.
[7] Q. Han, X. Gao and W. Wu, Study on Web Mining Algorithm Based on Usage Mining, 2010.
[8] Abdelhakim Herrouz, Chabane Khentout, Mahieddine Djoudi, "Overview of Web Content Mining Tools", the International Journal of Engineering and Science (IJES), 2013, 2, 6.
[9] J. Srivastava, R. Cooley, M. Deshpande and P. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, 2000, 1(2), 12–23.
[10] H. A. Edelstein, Pan for Gold in the Clickstream. Information Week, 2001, 77, 91.

★★★