IJCSC International Journal of Computer Science and Communication (ISSN: 0973-7391)
Vol. 3, No. 1, January-June, 2012, pp. 268-272

Web Mining- Research Issues and Challenges

Kalpana Gupta
Research Scholar, Department of Computer Science CMJ University, Shillong, INDIA

Abstract: As the number of web pages increases dramatically, the information overload become more and more severe when browsing and searching the WWW.Web is a collection of inter-related files on one or more web servers due to the wide range and availability of huge of heterogeneous data. The web has become versatile tool for almost all application today.Web mining is a new research discipline and it is a subdivision of data mining where data mining techniques are used for extracting information from the web servers. The web data includes web pages, web links, objects on the web and web logs.Web mining is used to understand the customer behaviour, evaluate a particular website based on the information which is stored in web log files.mining is evaluated by using data mining techniques,namely classification, clustering, and association rules. It has some beneficial areas or applications such as Electronic commerce, E-learning, E-government, E-policies, E-democracy, Electronic business, security, crime investigation and digital library. In this paper we have discuss the basic concept of web mining, classification, processes, issues and also analyzed the web mining research issues and challenges.

KEYWORDS: Web mining, Web content mining, Web structure mining, Web usage mining, Research issues, Challenges of web mining.

INTRODUCTION

In the late 1990s, low cost personal computers and an extensive, relatively easy to use Internet helped computers spread to the majority of households in may developed countries. Many of the activities for which people use the internet are long standing and well rooted in our social system. The web is an information of system distribution using the internet. The internet revolution is transforming to economics. In the world wide. No business sector, no company, will be remain untouched. Globally internet usershave grown from 39 million in 1995 to 315 million in 2000. The number is projected to grow to 716 million users in 2005. There were an estimated 2,459,846,518 internet users world in the February 2012, it is representing about 30.2 % of the population world wide, according to Internet world stats data updated in February 2012.

The World Wide Web is a rich source of information and continues to expand in size and complexity. Retrieving of the required web page on the web,efficiently and effectively,isbecoming a challenge. Whenever a user wants to search therelevant pages, he/she prefersthose relevant pages to be at hand. Relevant web page is one that providesthe same topic as the original page but it is not semantically identical to original page. Asthe Web is unstructured data repository,which delivers the bulk amount of information and also increases the complexity of dealing information from different perspective of knowledge seekers, business analysts and web service providers. Web involves three types of data; data on the WWW, the ger6web log data regarding the users who browsed the web pages and the web structure data.
Web mining can be easily executed with the help of other areas like Database (DB), Information retrieval (IR), Natural Language Processing (NLP), Machine Learning etc. Web mining can be defined as the discovery and analysis of useful information from the WWW data.

WEB MINING

Web mining process consists of four important steps, they are, resource finding, data selection and pre-processing, generalization and analysis.
Web mining is the Data Mining technique that automatically discovers or extracts the information

from web documents. It consists of following tasks:

1. Resource finding: It involves the task of retrieving intended web documents. It is the process by which we extract the data either from online or offline textresources available on web.
2. Information selection and preprocessing: It involves the automatic selection and pre processing of specific information from retrieved web resources. This process transforms the original retrieved data into information.
3. Generalization : It automatically discovers general patterns at individual web sites as well as across multiple sites. Data Mining techniques and machine learning are used in generalization
4. Analysis: It involves the validation and interpretation of the mined patterns. It plays an important role in pattern mining. A human plays an important role in information on knowledge discovery process on web

WEB MINING TEXONOMY

Web mining can be broadly classify in to three distinct categories, according to the kind of data to be mine.web content mining, web structure mining and web usage mining.This is illustrated in figure 1.

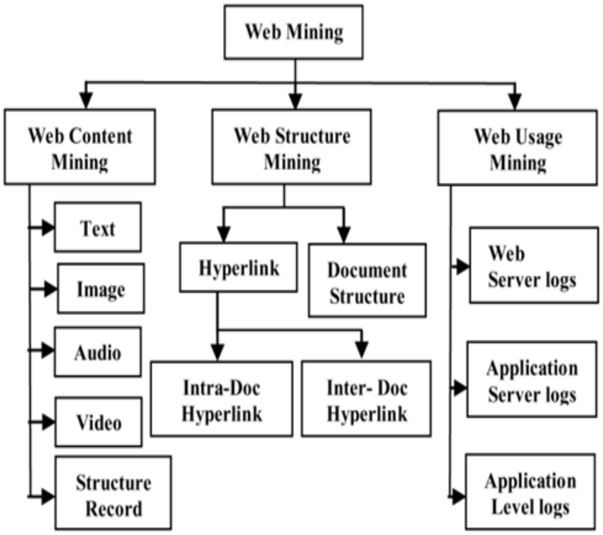

Figure 1. Classification of web mining

WEB CONTENT MINING

Web content mining is the process of retrieving the information from the web into more structured forms and indexing the information to retrieve quickly from web content or web documents. Web content mining includes the web documents which may consist of text, html, multimedia documents i.e., images, audio, video and sound etc. The search result mining contains the web search results. It may be a structure documents or unstructured documents. It is used to look at the information by search engine or web spiders i.e. Google, Yahoo. Web content mining used many algorithms and tools such as Genetic algorithm, Cluster Hierarchy Construction Algorithm (CHCA), Correlation algorithm Web Info Extractor (WIE) etc

WEB STRUCTURE MINING

Web structure mining is the study of data interconnected to the structure of a particular website. It consists of web graph which contains the web pages or web documents as nodes and hyperlinks as edges those are connecting between two related pages.Web structure is to extract some interesting web graph patterns like co-citation, social choice, complete bipartite graphs, etc. Web structure mining can be performed either at intra-page level or inter-page level. A hyperlink that connects to a different part of the same page is called intra-page hyperlink. A hyperlink that connects two different pages are called inter-page hyperlink which is structure level. Web structure mining used HITS (Hypertext Induced Topic Search) algorithm, Max flow- Min cut algorithm, ECLAT algorithm, and Page rank algorithm etc.

WEB USAGE MINING

Web usage mining is also called as web log mining which is used to analyze the behaviour of online users. Web usage mining is to extract the data which are stored in server access logs, referrer logs, agent logs and error logs. Web usage mining generally uses basic data mining algorithms such as association rule mining, sequential rule mining, clustering, and classification. It has several tools to analyze the behaviour of the user. They are web SIFT, web usage miner, INSITE, speed tracer, Archcollect, web miner, Web Quilt

RESEARCH ISSUES IN WEB MINING

The web is highly dynamic.Everyday lots of pages are added, updated and removed and it handles huge set of information hence there is an arrival of many number of problems or issues. Normally, web data is high dimensional, limited query interface, keyword oriented search and limited customization to individual users. Due to this, it is very difficult to find the relevant information from the web which may create new issues. Web measurement or web analytics are one of the significant challenges in web mining. The measurement factors are hits, page views, visits or user sessions and find the unique visitor regularly used to measure the user impact of various proposed changes. Large institutions and organizations archive usage data from the web sites. The main problem is that, detecting and/or preventing fraud activities. The web usage mining algorithms are more efficient and accurate. But there is a challenge that has to be taken into consideration. Web cleaning is the most important process but data cleaning becomes difficult when it comes to heterogeneous data.

Yang and Wu et al, (2006) discuss about the various issues to be addressed in data mining. The major issues includes

- Automated data cleaning
- Over fitting and Under fitting of data
- Over sampling of data
- Scaling up for high dimensional data
- Mining sequence and time series data
- Difficulty in finding relevant information
- Extracting new knowledge from the web
- Dealing with unbalanced data
- Unbalanced data
- Mining data streams
- Link and networks
- 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
- Proper organization of hardware and software to mine multi-terabyte data sets
- Limited customization, limited coverage, and limited query interface to individual users

CHALLENGES IN WEB MINING

1. Quality of keyword-based searches: The quality of keyword-based searches suffers from several inadequacies such as a search often returns many answers, especially if the keywords posed include words from popular categories such as sports, politics, or entertainment. It overloaded keyword semantics and it can return low-quality results. For example, depending on the context, an apple could be a fruit, juice, company or computer and a search can miss many highly related pages

2. Fraud and Threat Analysis: The main problem issue is that, are they ready for detecting and/or preventing fraud activities and can we completely remove the false positive and false negatives? The challenge is to find how we can gather knowledge directed data mining to eliminate false positives and false negatives. Another challenge of data mining is in real-time. The available tools of data mining have the ability to detect credit card violations and calling card violations. The research community should have a challenge to build a real time model. The challenge is necessary for many companies where they have interactions with up to millions of external parties. Details the subgroups of internal, insurance, credit card, and telecommunications fraud detection which is very concerned for both the researchers and particular organization.

3.Counter Terrorism: Privacy is a major challenge with respect to data mining for counter-terrorism. In this scenario, the challenge is to extract the structure and usage patterns or mine useful information form data mining but at the same time maintain privacy. Different efforts are under way for privacy preserving data.

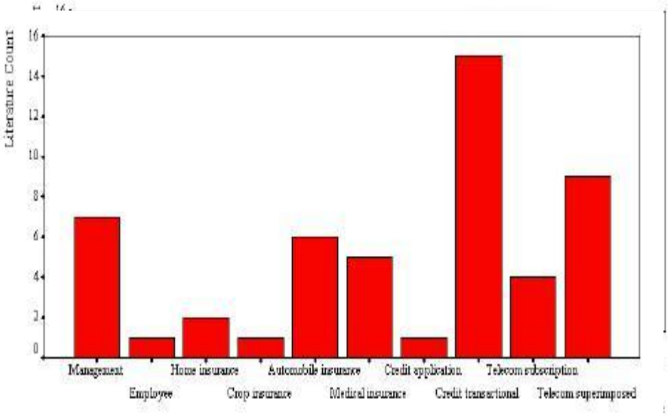

Figure 2. Bar chart of fraud types from 51 unique and published fraud detection papers [15]

There are various using techniques such as randomization, cover stories as well as multi party policy enforcement for privacy preserving data mining. That is while data mining could become a useful tool for counterterrorism, there are many challenges need to be addressed.

4. **Human activities feedback:** Web page authors provide links to ―authoritative Web pages and also traverse those Web pages they find most interesting or of highest quality. Unfortunately, while human activities and interests change over time, Web links may not be updated to reflect these trends. For example, significant events—such as the 2012 Olympic or the tsunami attack on Japan can change Web site access patterns dramatically, a change that Web linkages often fail to reflect. We have yet to use such human-traversal information for the dynamic, automatic adjustment of Web information services.

5. Effective of deep-Web Extraction: A research analysts estimated that searchable databases on the Web numbered more than 100,000. These databases provide high-quality, well-maintained information, but are not effectively accessible. Because current Web crawlers cannot query these databases, the data they contain remains invisible to traditional search engines. Conceptually, the deep Web provides an extremely large collection of autonomous and heterogeneous databases, each supporting specific query interfaces with different schema and query constraints. To effectively extract the deep Web, we must integrate these databases and implement efficient web mining approaches.

6.Security and Privacy: The monstrous maintenance of financial, demographic, behavioral, monetary, and other valuebased information for expository purposes may prompt the disintegration of common freedoms because of lost security and individual self-rule. From a protection and security point of view, the test is to guarantee that information subjects (i.e., people) have supportable control over their information, to anticipate abuse and mishandle by information controllers (i.e., enormous information holders and other outsiders), while saving information utility, i.e., the estimation of huge information for learning/designs revelation, advancement and financial development. The accompanying areas depict some applicable difficulties to security and protection with regards to enormous information.

CONCLUSION

This paper has discussed about the research issues and challenges in web mining and also provided detailed review about the basic concepts of web mining, web content mining, structure mining and web usage mining. Several open research issues and drawbacks which are exists in the current techniques are also discussed.this study and review would be helpful for researchers those who are doing their research in the domain of web mining.

REFERNCES

[1] J. Srivastava, R. Cooley, M. Deshpande and P. Tan, (2000) "Web usage mining: Discovery and applications of usage patterns from Web data", SIGKDD Explorations, Vol. 1, No. 2, pp. 12-23, 2000

[2] Mr. Dushyant B.Rathod, Dr.Samrat Khanna, "A Review on Emerging Trends of Web Mining and its Applications" ISSN: 2321-9939

[3] Joy Shalom Sona, Prof. Asha Ambhaikar" A Reconciling Website System to Enhance Efficiency with Web Mining Techniques" International Journal Of Scientific & Engineering Research Volume 3, Issue 2, February-2012 1 ISSN 2229-5518

[4] Jaideep Srivastava, "Web Mining: Accomplishments & Future Directions", University of Minnesota USA, srivasta@cs.umn.edu

[5] [1] Joy Shalom Sona, Prof. Asha Ambhaikar" A Reconciling Website System to Enhance Efficiency with Web Mining Techniques" International Journal Of Scientific & Engineering Research Volume 3, Issue 2, February-2012 1 ISSN 2229-5518.

[6] Sankar K. Pal, Varun Talwar, and Pabitra Mitra, "Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions", IEEE transactions on neural network, Vol. 13, No. 5, September 2002, pp.1163-1177

[7] Bhavani Thuraisingham, "Data Mining for CounterTerrorism", Chapter-3, MITRE Corporation, Burlington Road, Bedford, MA.

[8] Md. Zahid Hasan, Khawja Jakaria Ahmad Chisty and Nur-E-Zaman Ayshik, "Research Challenges in Web Data Mining", International Journal of Computer Science and Telecommunications Volume 3, Issue 7, July 2012

[9] Gerd Stummea, Andreas Hothoa, Bettina Berendtb,"SemanticWeb Mining: State of the art and futuredirections", Journal of Web Semantics, Vol. 4, Issue 2, June 2006, pp. 124–143.

[10]. Yu-Hui Tao, Tzung-Pei Hong, Yu-Ming Su,‖ Web usage mining with intentional browsing data‖ in international journal of Expert Systems with Applications 34 (2007) 1893–1904

[11]. Jiawei Han,Kevin,Chen-Chuan Chang "Data Mining for Web Intelligence" IEEE International Conference on Data Mining, 2002