International Journal of Information and Computation Technology (IJICT)

ISSN 0974-2239 Volume 3, Number 10 (2013)

Support System- Pioneering approach for Web Data Mining

Geeta Kataria[1], Surbhi Kaushik[2], Nidhi Narang[3]and Sunny Dahiya[4]

[1,2,3,4]Computer Science Department
Kurukshetra University
Sonepat, India

ABSTRACT

World Wide Web is an adequate source to find various kind of information. With an explosive growth of WWW, websites are playing an important role in providing an information and knowledge to the end users. With the broad use of internet, web data mining has gradually become the focus of current research on data mining. Web data mining i.e application of data mining techniques to extract knowledge from web content, structure and usage is the collection of technologies to extract valuable knowledge from the web. This paper provides a design & implementation of a research support system for web data mining to utilize useful information on the web.

Keywords- Web Data Mining, clustering, classification, association rules, web crawler, rank metrics

1.INTRODUCTION

Now days the World Wide Web is a popular and interactive medium to discriminate information. The web is huge, diverse, and dynamic and thus raises the scalability, multimedia data and temporal issues respectively. The need to understand large, complex, information-rich data sets is common to virtually all fields of business, science, and engineering. The ability to extract useful knowledge hidden in these data and to act on that knowledge is becoming increasingly important in today's competitive world. The entire process of applying a computer-based methodology for discovering and extracting knowledge from web documents is a web mining.

Two different approaches were taken in initially defining Web mining. First was a 'process-centric view', which defined Web mining as a sequence of tasks [H]. Second was a 'data-centric view', which defined Web mining in terms of the types of Web data that was being used in the mining process [G]. The second definition has become more acceptable, as is evident from the approach adopted in most recent papers [E, M, A] that have addressed the issue. In this paper we follow the data-centric view of Web mining which is defined as follows,

Web mining is the application of data mining techniques to extract knowledge from Web data, i.e. Web Content, Web Structure and Web Usage data. The attention paid to Web mining, in research, software industry, and Web-based organizations, has led to the accumulation of a lot of experiences.

The rest of this paper is organized as follows: In Section 2 we provide taxonomy of Web mining, in Section 3 we discuss research support system framework overview for data mining, and in section 4 we describe how to retrieve information ,in section 5 we describe successful applications of Web mining techniques. In Section 6 we conclude the paper and present some directions for future research.

2. WEB MINING TAXONOMY

Web Mining can be broadly divided into three distinct categories based on [A] and [D]. We provide a brief overview of the three categories. A figure depicting the taxonomy is shown in Fig.1.

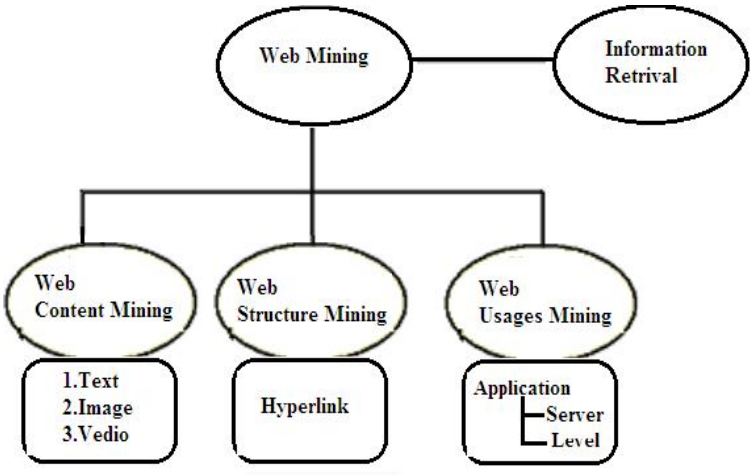


Fig.1 Web Mining Taxonomy

2.1 Web Content Mining

Web Content Mining is the process of extracting 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, or structured records such as lists and tables. Web content mining issues in term of Information Retrieval (IR) and Database (DB) view verses data representation, method and application categories is discuss and summarized in [M]. While extracting the knowledge from images - in the fields of image processing and computer vision -the application of these techniques to Web content mining has not been very rapid.

2.2 *Web Structure Mining*

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages it is the process of discovering structure information from the Web.

2.2.1 *Hyperlinks***:** A Hyperlink is a structural unit that connects a location in a Web page to different location, either within the same Web page or on a different Web page. A hyperlink that connects to a different part of the same page is called an *Intra-Document Hyperlin*k, and a hyperlink that connects two different pages is called an *Inter*-Document Hyperlink. There has been a significant body of work on hyperlink analysis provide an up-to-date survey[K].

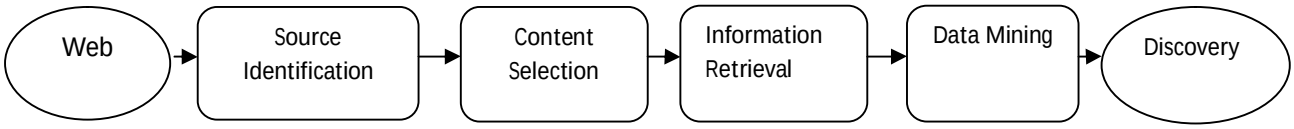
2.2.2 *Document Structure*: In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents [N, P].

2.3 *Web Usage Mining*

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from web data, in order to understand and better serve the needs of web-based applications. Some of typical data mining methods used to mine the usage data after the data have been pre-processed in desired form. Modification of typical data mining is used composite associative rule, extraction of traditional sequence discovery algorithm, web usage using graph representation, is discussed in and can referred from [M].

3. RESEARCH SUPPORT SYSTEM

In order to explore web data, we construct a research support system framework for web data mining, as shown in Fig 2, consisting of four phases: source identification, content selection, information retrieval and data mining. In the first phase, proper web sites should be chosen according to research needs.



This includes identifying avail-ability, relevance and importance of web sites. Key words searching by using search engine can be used to find appropriate web sites.

Figure.2 Research support system framework for web data mining

After finding all web sites identified by the first phase, the second phase is to select appropriate contents on those web sites, such as documentation, newsgroups, forums, mailing lists, etc. Usually, a web site contains many web pages, including relevant and irrelevant information. This phase is important because it decides which web information should be extracted. The selection of web pages is based on re-search purpose and a researcher's experience.

In the information retrieval phase, a crawler is designed to automatically extract information selected during the selection phase. Specific tools and techniques are employed to effectively retrieve useful knowledge/information from web sources. Additional effort may be required for dynamic content retrieval and specific data sources such as newsgroup, forum, etc. The final phase is to conduct data mining on extracted web data. It includes preparing data for analysis. An extracted web page may contain missing data, extraneous data, wrong format and unnecessary characters. Furthermore, some data should be processed in order to protect privacy. Advanced data mining techniques are employed here to help analyzing data.

4. INFORMATION RETRIEVAL

Information retrieval is used to provide access to data on the web. This implies a web mining research support system should be able to search for and retrieve specific contents on the web efficiently and effectively. There are two major categories of searching tools on the Web: directories (Yahoo, Netscape, etc.) and search engines (Lycos, Google, etc.). It is hard to use directories with the increase of web sites. Search engines cannot meet every search requirement.

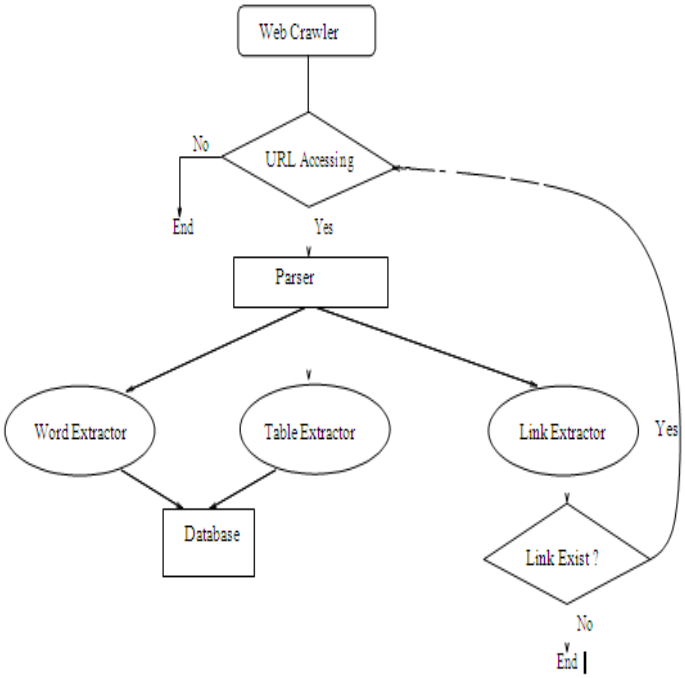


Figure.3 Web Crawler

In our system, a web crawler based on advanced tools and techniques is developed to help find useful information from web resources. Web crawlers are also called spiders, robots, worms, etc. A web crawler is a program which automatically traverses web sites, downloads documents and follows links to other pages [C]. It keeps a copy of all visited pages for later uses. Many web search engines use web crawlers to create entries for indexing. They can also be used in other possible applications such as page validation, structural analysis and visualization, update notification, mirroring and personal web assistants/agents etc.[B]. Search engines are not adequate for web mining for a research project. It is necessary to design a web crawler which includes methods to find and gather the research related information from the web.

Although different research projects have different web information which leads to different web crawlers, those crawlers still have some common designs, as shown in Figure 3. They can be implemented by Java, Perl, Python, etc.

A web crawler should start with a URL, identify other links on the HTML page, visit those links, extract information from web pages and store information into databases. Thus, a web crawler consists of a URL access method, a web page parser with some extractors, and databases. The access function of a web crawler should prevent repeatedly accessing the same web address and should identify dead links. The parser recognizes start tags, end tags, text and comments. The databases provide storage for extracted web information.

The key component of a web crawler is the parser, which includes a word extractor, a table extractor and a link extractor. The word extractor is used to extract word information. It should provide string checking functions. Tables are used commonly in web pages to align information. A table extractor identifies the location of the data in a table. A link extractor retrieves links contained in a web page. There are two types of links – absolute links and relative links. An absolute link gives the full address of a web page, while a relative link needs to be converted to a full address by adding a prefix.

5. DATA MINING TECHNOLOGY

To facilitate web mining, the following data mining algorithms can be applied to find patterns and trends in the data collected from the web: clustering, classification, association rules.

6.1 Association Rules

Association rules mining tries to find interesting association or correlation relationship among a large set of data items. A typical example of association rules mining is the market basket analysis. Association rules mining can also be applied to predict web access patterns for personalization. For example, we may discover that 80% of people who access page A and page B also access page C. Page C might not have a direct link from either page A or page B. The information discovered might be used to create a link to page C from page A or page B. One example of this application is amazon.com. We often see something like ”customers who buy this book also buy book A”. The association rules mining can be applied to web data to explore the behavior of web users and find patterns of their behaviors.

6.2 Classification

Classification is the task of mapping a data item into one of several predefined classes.The goal of classification is to predict which of several classes a case (or an observation) belongs to. Each case consists of n attributes, one of which is the target attribute, all others are predictor attributes. Each of the target attribute’s value is a class to be predicted based on the n _ 1 predictor attributes. Classification is a two-step process. First, a classification model is built based on training data set. Second, the model is applied to new data for classification. In the middle of the two steps, some other steps might be taken, such as lift computation. Lift computation is a way of verifying whether a classification model is valuable. A value larger than 1 is normally good. Classification models can be applied on the web to make business decisions. Applications include classifying email messages as junk mails, detecting credit card fraud, network intrusion detection, etc.

6.3 Clustering

Clustering is used to find natural groupings of data. These natural groupings are clusters. A cluster is a collection of data that are similar to one another. A good clustering algorithm produces clusters such that inter-cluster similarity is low and intra-cluster similarity is high.

6. CONCLUSIONS and FUTURE WORK

Web data mining is a new research field; it is different from the traditional data mining. Due to the Web data is an unstructured data, it makes the data mining become very difficult. So, in this paper we have discussed research support system that is used to extract useful information and describes its procedures. It then discusses implementing techniques on web data extraction and analysis. This work is an exploratory study of web data retrieval and data mining on web data. The actual interesting discoveries are still in progress.

7. REFERENCES

A S. K. Madria, S. S. Bhowmick, W. K. Ng,and E. P. Lim**(1999),** *Research Issues in Web Mining* , in proceding of *data mining and knowledge discovery, 1st International conference, DaWk 99, pp 303-312.*

A S.K. Madria, S.S. Bhowmick, W.K Ng, and E.P Lim**(1999).** *Research Issues in Web Data Mining. In Data Warehousing and Knowledge Discovery, pages 303–312.*

B Francis Crimmins **(2001).** Web crawler review. http://dev.funnelback.com/crawler-review.html.

C M. Koster**(1999).** The web robots pages. http://info.webcrawler.com/mak/projects/robots/robots.html.

D J. Borges abd M. Lelene**(1999),** *"Data Mining of Use Navigation Pattern", in proceeding of WEBKDD 99,Workshop on web usage analysis and user profiling, CA USA, pp 31-36.*

E J. Borges and M. Levene**(1998).** *Association Rules in Hypertext Databases in Knowledge Discovery and Data Mining, pages 149–153.*

F S. Brin and L. Page**(1998).** *The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30(1-7):107–117.*

G R. Cooley, J. Srivastava, and B. Mobasher**(1997).** *Web mining: Information and pattern discovery on the world `wide web. In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97).*

H O. Etzioni**(1996).** *The World-Wide Web: Quagmire or Gold Mine? Communications of the ACM, 39(11):65–68.*

I Jain A.K., Murty M.N., Flyn, P.J. **(1999),** *"Data Clustering: A Review", ACM Computing Surveys, Vol. 31, No. 3, pp. 264 - 323.*

J Ester, M., Kriegel, H.-P., Sander, J., and Xu X**.(1996),** *"A Density-based Algorithm for Discovering Clusters in Large Spatial Data Sets with Noise",. Proc.2nd Int. Conf. on Knowledge Discovery and Data Mining. Portland, pp. 226–231.*

K P. Desikan, J. Srivastava, V. Kumar, and P.N. Tan(**2002**). *Hyperlink AnalysisTechniques &Applications. Technical Report 2002-152, Army High Performance Computing Research Center.*

L J.M. Kleinberg(**1999**).*Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604–632.*

M R. Kosala and H. Blockeel**(2000**). *Web mining research*: *A survey. SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery& Data Mining, AC*M, 2.

N C.H. Moh, E.P. Lim, and W.K. Ng**(2000).** *DTD-Miner: A Tool for Mining DTD from XML Documents. WECWIS.*

O L. Page, S. Brin, R. Motwani, and T. Winograd**(1998**). The page rank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project.

P K. Wang and H. Liu**(1998).** *Discovering Typical Structures of Documents: A Road Map Approach. In 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pages 146–154.

Q Google Inc. *http://www.google.com.*