ISSN:0975-9646

The STUDY on KNOWLEDGE DISCOVERY from WEB DATA

Yogish H K[1] ,Dr. G T Raju2,Deepa Yogish[3]

[1]Department of Computer Science and Engineering,REVA Institute of Technology and Management, Yelahanka. Bangalore-560064,Karnataka, India.(Research Scholar - Bharathiar University, Coimbatore-641046)

[2]Department of Computer Science and Engineering,RNS Institute of Technology, Bangalore -560061.

[3]Department of Computer Science and Engineering,Don Bosco Institute of Technology, Bangalore -560074.

Abstract - **The World Wide Web serves as huge, widely distributed global information service centre for news, advertisements, consumer information, financial management, education, government, e-commerce and many other information services. The web also contains a rich and dynamic collection of hyperlink information and web page access and usage information, providing rich sources of for data mining. However, based on the following observations the web also poses great challenges for effective resource and knowledge discovery. The Web usage mining is the area of data mining which deals with the discovery and analysis of usage patterns from Web data, specifically web logs, in order to improve web based applications. Web usage mining consists of three phases, preprocessing, pattern discovery, and pattern analysis. After the completion of these three phases the user can find the required usage patterns and use this information for the specific needs.**

***Keywords* – Web data, data mining, knowledge discovery, text mining**.

I. Introduction

The web is the massive collection of documents stored in one or more web servers. These documents are unorganized. The web mining or knowledge discovery from web is the process of discovering potentially useful and previously unknown information or knowledge from the web data.

The web data is:

- Web content data: such as text, images, records, audio, video and metadata.
- Web structure data: such as hyperlinks, tags.
- Web usage data: such as server logs, application server logs, browsers logs, user's profiles, user's queries, book marked data, mouse clicks and scrolls, registration data, cookies, user sessions or transactions data.

II. Web Challenges [1, 2]

The web seems to be huge for effective data warehousing and data mining: The size of the web is in the order of hundreds of terabytes and still growing rapidly. Many organizations and societies place most of their public accessible information on the web. The complexity of web pages is far greater than that of any traditional text document collection. Web pages lack a unifying structure.

The web is considered a huge digital library: However the tremendous numbers of documents in this library are arranged according to any particular sorted order. There is no index by category or by title, author, cover page, table of contents and so on. It can be very challenging to search for the information you desire in such a library.

The web is highly dynamic information source: Not only does the web grow rapidly, but its information is also constantly updated. News, stack markets, weather, sports, shopping, company advertisements and numerous other web pages are updated regularly on the web.

Only a small portion of information on the web is truly relevant or useful: It is said to that 99% of the web information is useless to 99%of web users. How can, the portion of the web that is truly relevant to your interest be determined?

These challenges have promoted research into efficient and effective discovery and use of resources on the Internet.

III. Web Usages

We make use of web in several ways; we interact with the web for the following purposes.

A. Finding relevant information

We either browse or use the search service when we want to find specific information on the web. We usually specify a simple keyword query and the response from the web search engine is a list of pages, ranked based on their similarity to the query.

However today's a search tools have the following problems:

a. *Low precision*

This is due to the irrelevance of many of the search results; we may get many pages of information which are not really relevant to our query.

b. *Low recall*

This is due to the inability to index all the information available on the web, because some of the relevant pages are not properly indexed, we may not get those pages through any of the search engines.

B. *Discovering new knowledge from the web*

From the collected web data, extract potentially useful and previously unknown Information or knowledge.

C. Personalized web synthesis

We may wish to synthesize a web page for different individuals from the available set of web pages. Individuals have their own preferences in the style of the contents and presentations while interacting with the web.

D LEARNING ABOUT INDIVIDUAL USERS OR CUSTOMERS

It is about knowing what the customers do and want. I.e. customizing the information to the intended customers or even personalizing it to individual user, problems related to effective website design and management, problem related to marketing etc.

Web mining techniques provide a set of techniques that can be used to solve the above problems.

Web mining looks for three data mining terms:

1. Clustering: finding natural groupings of users, pages etc.
2. Associations: which URL tend to be requested together
3. Sequential analysis: the order in which URLs tend to be accessed.

IV. WEB TAXONOMY

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined:

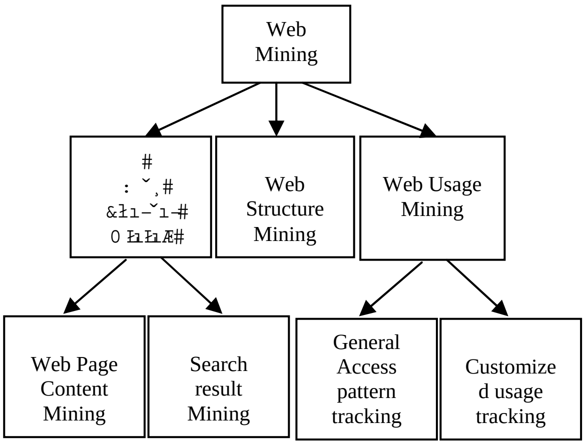

Figure 1: Taxonomy of Web Mining

A. WEB CONTENT MINING

Describes the discovery of useful information from web contents / data / documents.

The web content consists of several types of data such as textual, image, audio, metadata and hyper links. The web content mining concentrates on the text or hyper text contents.

The web content data may be: [3,4,5]

1. Unstructured data such as free test
2. Semi structured data such as HTML documents.
3. More structured data such as data in the tables or database generated HTML pages.

The techniques of text mining can be used for web content mining. Traditional search engines such as Lycos, Alta Vista, WebCrawler, ALIWEB, Meta Crawler, and others provide some comfort to users, but do not generally provide structural information nor categorize, filter or interpret documents.

Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting 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.

B. WEB USAGE MINING

Is also known as weblog mining. Performs mining on web usage data or web logs. Weblog is a listing of page reference data, some time it is referred as click streams data because each entry corresponds to mouse click. It deals with studying the data generated by the web surfer's sessions. Or Discover the usage pattern from the secondary data derived from the interaction of the users while surfing the web[1].

The secondary data includes the data from the web server access logs, application server logs, browsers logs, user's profiles, user's queries, book marked data, mouse clicks and scrolls, registration data, cookies, user sessions or transactions data. This data can be accumulated by the web server. Analysis of web access logs of different web sites can facilitate and understanding of the user behavior.

The logs can be examined from either server or client perspective. When evaluated from a server perspective, mining uncovers informs about a sites where the service resides, it can be used to improved the design of the sites. By evaluating client's sequence of clicks information about the users or group of users or detected. This could be used to perform pre-fetching and caching of pages.

For Example: The web master at ABC [3] corporation learns that a high percentage of users have the following patterns of reference to pages :(A, B, A, C). This means that user access page A, then page B, then back to page A and finally to page C. Based on this observation he determines that a link is needed directly to page C from B. He then adds this link.

1) *GENERAL ACCESS PATTERN TRACKING*
   This is to learn user navigation patterns (impersonalized). This analyses the web logs to understand access pattern and trends.
2) *CUSTOMIZED USAGE TRACKING*
   This is to learn user profile (personalized). It analyses the individual trends i.e. is analyses access pattern of each user at a time. Its purpose is to customize web sites to users.

V 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. Web Structure Mining can be regarded as the process of discovering structure information from the Web. This type of mining can be further divided into two kinds based on the kind of structural data used.

A. Hyperlinks[1]

A Hyperlink is a structural unit that connects a Web page to different location, either within the same Web page or to a different Web page. A hyperlink that connects to a different part of the same page is called an *Intra-Document Hyperlink*, and a hyperlink that connects two different pages is called an *Inter-Document Hyperlink.*

B. Document Structure [1]

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.

We can view any collection, V of hyper linked pages as a directed graph:

$$G = (V, E).$$

The nodes corresponds the pages and a directed edge $(p,q) \subset E$ indicates the presence of a link from p to q.

We say that out degree of a node *p* is the number of nodes to which it has links and in degree of node *p* is the number of nodes that have links to it. The following algorithms are proposed to model web topology such as HITS, page rank and CLEVER.

C. Page Rank[1]

The page rank technique was designed to improve the effectiveness and efficiencies of search engines. Is used to measures the importance of a page and prioritize pages returned from a traditional search engine using key word searching. The page rank value for a page is calculated based on the number of pages that point to it. Page rank is defined as follows: we assume page A has pages p_1, $p_{2,}$,…,p_n which points to it. The parameter'd' is a damping factor which can be set between 0 and 1 and it is usually 0.85. Out degree(A) denotes the number of links going out of page A. the page rank of a page A is given as follows:

$$PR(A)=(1-d) + d\left(\sum_{i=1}^{n} \frac{PR(p_i)}{outdegree(p_i)}\right)$$

Page rank is calculated using simple iterative algorithms.

VI DATA PREPROCESSING FOR MINING

Web data is collected in various ways[7], each mechanism collecting attributes relevant for its purpose. There is a need to preprocess the data to make it easier to mine for knowledge, specifically; we believe that issues such as instrumentation and data collection, data integration and transaction identification need to be addressed.

Clearly improved data quality can improve the quality of any analysis on it. A problem in the Web domain is the inherent conflict between the analysis needs of the analysts, who want more detailed usage data collected, and the privacy needs of users, who want as little data collected as possible. This has lead to the development of cookie files on one side and cache busting on the other, the emerging OPS standard on collecting profile data may be a compromise on what can and will be collected. However, it is not clear how much compliance to this can be expected. Hence, there will be a continual need to develop better instrumentation and data collection techniques, based on whatever is possible and allowable at any point in time.

Web usage data collected in various logs is at a very fine granularity, Therefore, while it has the advantage of being extremely general and fairly detailed, it also has the corresponding drawback that it cannot be analyzed directly, since the analysis may start focusing on micro trends rather than on the macro trends, On the other hand, the issue of whether a trend is micro or macro depends on the purpose of a specific analysis.

Hence, we believe there is a need to group individual data collection events into groups, called Web transactions, before feeding it to the mining system.

VII THE MINING PROCESS

The key component of Web mining is the mining process itself. The Web mining has adapted techniques from the field of data mining, databases, and information retrieval, as well as developing some techniques of its own.

Web mining studies reported to data have mined for association rules, temporal sequences, clusters, and path expressions. As the manner in which the Web is used continues to expand, there is a continual need to figure out new kinds of knowledge about user behavior that needs to be mined.

The quality of a mining algorithm can be measured both in terms of how effective it is in mining for knowledge and how efficient it is in computational terms. There will always be a need to improve the performance of mining algorithms along both these dimensions.

The data collection on the Web is incremental in nature. Hence, there is a need to develop mining algorithms that take as input the existing data mined knowledge, and the new data, and develop a new model in an efficient manner.

The data collection on the Web is also distributed by its very nature. If all the data were to be integrated before mining, a lot of valuable information could be extracted.

VIII ANALYSIS OF MINED KNOWLEDGE

The output of knowledge mining algorithms is often not in a form suitable for direct human consumption, and hence there is a need to develop techniques and tools for helping an analyst better assimilate it. Issues that need to be addressed in this area include usage analysis tools and interpretation of mined knowledge.

There is a need to develop tools which incorporate statistical methods, visualization, and human factors to help better understand the mined knowledge.

In general one of the open issues in Web mining in particular, is the creation of intelligent tools that can assist in the interpretation of mined knowledge. Clearly, these tools need to have specific knowledge about the particular problem domain to do any more than filtering based on statistical attributes of the discovered rules or patterns. In Web mining, for example, intelligent agents could be developed that based on discovered access patterns, the topology of the Web locality, and certain heuristics derived from user behavior models, could give recommendations about changing the physical link structure of a particular site.

IX CONCLUSION

The term Web mining has been used to refer to techniques that encompass a broad range of issues However□ while meaningful and attractive. This very broadness has caused Web mining to mean different things to different people and there is a need to develop a common vocabulary. Towards this goal we proposed a definition of Web mining and developed taxonomy of the various ongoing efforts related to it. Next, we presented a survey of the research in this area. We provided a detailed survey of the efforts in this area, even though the survey is short because of the area's newness. This paper is useful for researcher exclusively for doing research on web mining.

REFERENCES

[1] Arun k Pujari, "*Data Mining Techniques*", University press, edition 2001.

[2] Jaiwei Han, Michelinne Kamber, "*Data Mining: Concepts and Techniques* ".

[3] Margaret H.Dunham, "*Data Mining:Introductory and advanced topics*".

[4] Srivastava, J., Cooley, R., Deshpande, M., and Tan, P., *Web usage mining: Discovery and applications of usage patterns from Web data, SIGKDD Explorations*, Vol. 1(2), 12-23, 2000.

[5] *Knowledge Discovery from Web Usage Data: Extraction of Sequential Patterns through ART1 Neural Network based Clustering Algorithm*, International Conference on Computational Intelligence and Multimedia Applications 2007

[6] Jaideep Srivastava, Prasanna Desikan , Vipin Kumar .,*Web Mining – Accomplishments & Future Directions.*

[7] "*Mining Web logs for Prediction in Prefetching and Caching*" - Third 2008 International Conferences on Convergence and Hybrid Information Technology.

AUTHORS' BIBLIOGRAPHY

YOGISH H. K received his Bachelor's Degree in computer Science and Engineering from PES College of Engineering, Mandya, Mysore University, Karnataka, India during the year 1988 and M. Tech in Computer Engineering from Sri Jaya Chamarajendra College of Engineering Mysore, Karnataka, India during the year 2004. Currently pursing PhD degree in Bharathiar University, Coimbatore. Am having total 12 years of Industry and teaching experience. My areas of interests are Data Warehouse, multimedia, Databases and Operating Systems. I have published and presented papers in journals, International and national level conferences and an author of Two Text Books.

Dr. G T Raju received his Bachelor's Degree in Computer Science and Engineering from Kalpataru Institute of technology, Tiptur, Bangalore University, Karnataka, India, during the year 1992 and M. E in Computer Science and Engineering from B.M.S College of Engineering, Bangalore, Bangalore University, Karnataka, India during the year 1995 and Doctorate of Philosophy Ph.D. in year 2008 in Computer science from Visveswaraya Technological University, Belgaum, Karnataka; He is having 18 years of Experience. He has visited overseas to various universities. His area of interests is Data Mining, data Warehousing, Image Processing and Databases, Artificial Intelligence and Computer Graphics. He has published and presented papers in journals, international and national level Conferences and published a text book.

Deepa Yogish received his Bachelor's Degree in Electronics and Communication Engineering Visveswaraya Technological University, Belgaum, Karnataka, India during the year 1988 and M. Tech in Computer Science and Engineering from Visveswaraya Technological University, Belgaum Karnataka during the year 2009. Having total 6 years of teaching experience. My areas of interests are Data Warehouse, Databases and Operating Systems. I have published and presented three papers in national conferences.