Volume 5, Issue 9, September 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Survey on Web Mining and Its Techniques

[1]P. Menaka MCA., M.Phil, [2]A. Prathimadevi

[1]Assistant Professor, Department of Information Technology, Dr.N.G.P Arts and Science College, Coimbatore, India

[2]Research Scholar, Department of Computer Science, Dr.N.G.P Arts and Science College, Coimbatore, India

***Abstract:** Nowadays most of the people rely on internet. At the same time internet has many information. It should provide related information for each user query. Web mining is used to extract information based on the user query from the large collection of data available in web. It is concerned mainly with its content, structure and usage. Web usage mining is extracting information based on the user log, frequently accessed paths. Web content mining is used to fetch information from the web documents. Web structure mining usually use graph theory to extract the web site structure through which they provide better search results for the user. This paper also reports the summary of various techniques of web mining approached from the following angles like Feature Extraction, Transformation and Representation and Data Mining Techniques in various application domains*

***Keywords:** Web Mining, Web Usage Mining, Web Structure Mining, Web Content Mining, Graph theory.*

I. INTRODUCTION

The web is very enormous, diverse, flexible, and dynamic. The World Wide Web continues to grow both in the huge volume of traffic and the size and complexity of Web sites. With the increasing growth of information available in net. It is difficult to identify the relevant information present in the web. Meanwhile much information is unstructured. It is necessary to use automated tool for obtaining the necessary information from the huge collection of information.

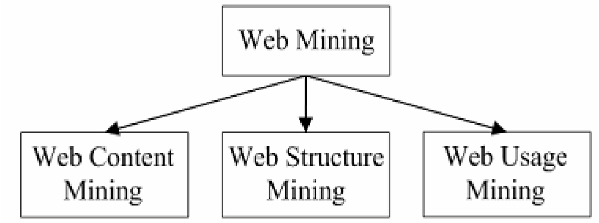


Figure 1: Taxonomy of Web Mining

Web Mining is used to extract information from the raw unstructured data. The emerging field of web mining aims at finding and extracting relevant information that is hidden in Web related data, in particular in text documents published on the Web. Web mining is performed in three ways they are 1) web usage mining 2) web content mining 3) web structure mining.Web usage mining provides the support for the web site design, providing personalization server and other business making decision, etc. Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web document text mining, resource discovery based on concepts indexing or agent; based technology may also fall in this category. Web structure mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web. Finally, web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in web access logs. In order to better serve for the users, web mining applies the data mining, the artificial intelligence and the chart technology and so on to the web data and traces users visiting characteristics, and then extracts the users' using pattern. It has quickly become one of the most important areas in Computer and Information Sciences because of its direct applications in e-commerce, CRM, Web analytics, information retrieval and filtering, and Web information systems.

II. WEB MINING

Web mining is the use of data mining techniques to automatically extract data from the web. Web mining has various sub tasks they are 1) Resource finding 2) Information Selection and pre-processing 3) Generalisation 4) Analysis

Resource Finding is the task of retrieving intended Web documents. Information Selection and pre-processing is automatically selecting and pre-processing specific information from retrieved web resources. Generalization is automatically selecting and pre-processing specific information from retrieved wed resources. Analysis is validation and interpretation of mined patterns.

III. WEB USAGE MINING

Web usage mining is used to extract information based on the user log. Web Usage mining is the process of applying data mining techniques to the discovery of usage patterns from Web data, targeted towards various applications.

The usage data collected at the different sources will represent the navigation patterns of different segments of the overall Web traffic, ranging from single-user, single-site browsing behavior to multi-user, multi-site access patterns.

3.1 CONCEPT OF WEB USAGE MINING

Discovery of meaningful patterns from data generated by client-server transactions on one or more web servers.
Typical Sources of Data:

1. Automatically generated data stored in server access logs, referrer logs, agent logs, and client-side cookies.
2. E-commerce and product-oriented user events (e.g. shopping cart changes, ad or product click-through, etc).
3. User profiles and/or user ratings.
4. Meta-data, page attributes page content, site structure.

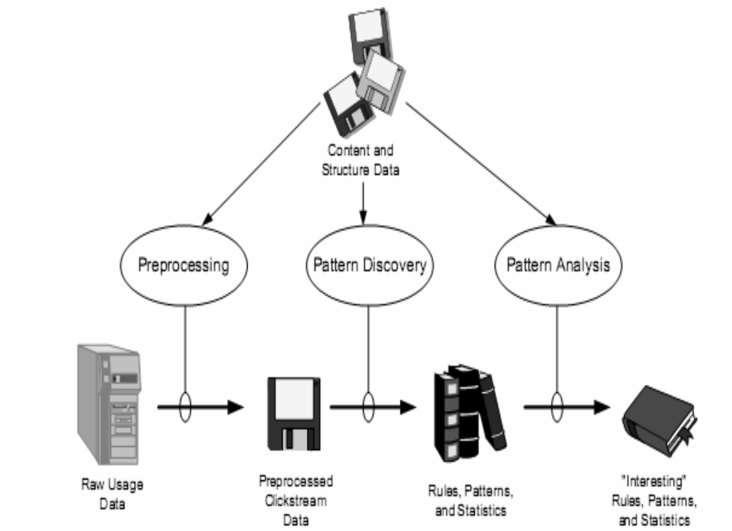
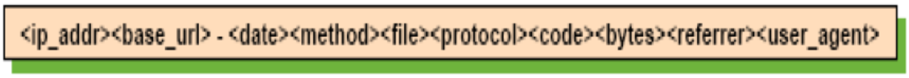

Content and Structure Data
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Usage Data
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

Figure 3: Web Usage Mining

IV. WEB LOG FORMAT

A web server log file contains requests made to the web server, recorded in chronological order. The most popular log file formats are the Common Log Format (CLF) and the extended CLF. A common log format file is created by the web server to keep track of the requests that occur on a web site.



<ip_addr><base_url> - <date><method><file><protocol><code><bytes><referrer><user_agent>

Figure 4: Web log Format

V. APPROACHES OF WEB USAGE MINING

A) Data collection:

Data collection is the first step of web usage mining, the data authenticity and integrality will directly affect the following works smoothly carrying on and the final recommendation of characteristic service's quality. Therefore it must use scientific, reasonable and advanced technology to gather various data. At present, towards web usage mining technology, the main data origin has three kinds: server data, client data and middle data

B) Data preprocessing:

Some databases are insufficient, inconsistent and including noise. The data pretreatment is to carry on a unification transformation to those databases. The result is that the database will to become integrate and consistent.

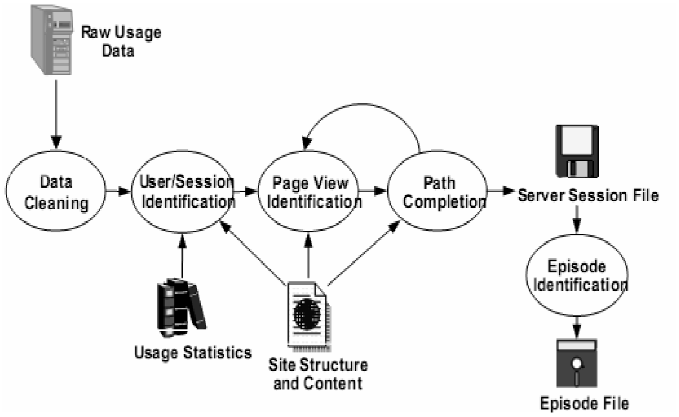


Figure 5: Preprocessing of Web Usage Data

C) Knowledge Discovery:

Use statistical method to carry on the analysis and mine the pretreated data. We may discover the user or the user community's interests then construct interest model. At present the usually used machine learning methods mainly have clustering, classifying, the relation discovery and the order model discovery. Each method has its own excellence and shortcomings, but the quite effective method mainly is classifying and clustering at the present.

D) Pattern analysis:

Challenges of Pattern Analysis are to filter uninteresting information and to visualize and interpret the interesting patterns to the user. First delete the less significance rules or models from the interested model storehouse; Next use technology of OLAP and so on to carry on the comprehensive mining and analysis; Once more, let discovered data or knowledge be visible; Finally, provide the characteristic service to the electronic commerce website.

VI. WEB CONTENT MINING

The Web content mining refers to the discovery of useful information from web contents which include text, image, audio, video, etc. The mining of link structure aims at developing techniques to take advantage of the collective judgment of web page quality which is available in the form of hyperlinks that is web structure mining. It includes extraction of structured data/information from web pages, identification, similarity and integration of data's with similar meaning, view extraction from online sources, and concept hierarchy, knowledge incorporation.

6.1 Techniques in Web Content Mining:

A) Classification of Multimedia Content and Websites

In order to retrieve relevant knowledge a system has to analyze web content first. The Classification of web objects offers an automatic way to decide the relevance of web objects. Since websites are usually represented by multiple pages, classifying website on top of web pages classification demands new algorithms

B) Focused Crawling

A focused web crawler takes a set of well-selected web pages exemplifying the user interest. The focused crawler starts from the given pages and recursively explores the linked web pages. While the crawlers perform a breadth-first search of the whole web, a focused crawler explores only a small portion of the web using a best-first search guided by the user interest. Crawling for retrieving multimedia content in the web, instead of plain HTML documents.

C) Clustering Web Objects

Focused Crawling retrieves large numbers of relevant data.In order to offer fast and more specific access to the query results, clustering is an established method to group the retrieved information to achieve better understanding. If the query results are websites or combined objects like images and their text descriptions, algorithm are needed to handle these combined data types to find meaningful clustering.

D) Wrapper Induction

A wrapper is a piece of software that enables a semi structured Web source to be queried as if it were a database Given a set of manually labeled pages, a machine learning method is applied to learn extraction rules or patterns.

E) Automatic Data Extraction

Given a set of positive pages, generate extraction patterns. Given only a single page with multiple data records, generate extraction patterns.

VII. WEB STRUCTURE MINING

Web structure mining focuses on the hyperlink structure of the Web. The different objects are linked in some way. Simply applying the traditional processes and assuming that the events are independent can lead to wrong conclusions. The challenge for Web structure mining is to deal with the structure of the hyperlinks within the Web itself. Link analysis is an old area of research. However, with the growing interest in Web mining, the research of structure analysis had increased and these efforts had resulted in a newly emerging research area called Link Mining, which is located at the intersection of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining. There is a potentially wide range of application areas for this new area of research, including Internet.

7.1 TASKS IN WEB STRUCTURE MINING

A) Link-based Classification

Link-based classification is the most recent upgrade of a classic data mining task to linked domains. The task is to focus on the prediction of the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags and other possible attributes found on the web page.

B) Link-based Cluster Analysis

The goal in cluster analysis is to find naturally occurring sub-classes. The data is segmented into groups, where similar objects are grouped together, and dissimilar objects are grouped into different groups. Different than the previous task, link-based cluster analysis is unsupervised and can be used to discover hidden patterns from data.

C) Link Type

There are a wide range of tasks concerning the prediction of the existence of links, such as predicting the type of link between two entities, or predicting the purpose of a link.

D) Link Strength

Links could be associated with weights.Each link should have unique weight

E) Link Cardinality

The main task here is to predict the number of links between objects.

VIII. CONCLUSION

In this paper we survey the research area of Web mining, focusing on the category of Web structure mining. We had introduced ***Web mining***. Later in the paper when we had discussed Web structure mining, and introduced Link mining, as well as block-level link mining issues. We had also reviewed two popular algorithms to have an idea about their application and effectiveness. Since this is a huge area, and there a lot of work to do, we hope this paper could be a useful starting point for identifying opportunities for further research.

REFERENCES

[1] http://maya.cs.depaul.edu/~classes/ect584/papers/srivastava.pdf

[2] Web Structure Mining: An Introduction,Miguel Gomes da Costa Júnior Zhiguo Gong Department of Computer and information Science,Faculty of Science and Technology, University of Macau published in international conference on information acquisition.

[3] O. Etzioni. The world wide web: Quagmire or gold mine. *Communications of the ACM*, 39(11):65-68, 1996.

[4] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey, *ACM SIGKDD Explorations Newsletter*, June 2000, Volume 2 Issue 1.

[5] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pag-Ning Tan,Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, *ACM SIGKDD Explorations Newsletter*, January 2000, Volume 1 Issue 2.

[6] Jidong Wang, Zheng Chen, Li Tao, Wei-Ying Ma, Liu Wenyin, Ranking User's Relevance to a Topic through Link Analysis on Web Logs, *WIDM' 02*, November 2002.

[7] A. A. Barfourosh, H.R. Motahary Nezhad, M. L. Anderson, D. Perlis, Information Retrieval on the World Wide Web and Active Logic: A Survey and Problem Definition, 2002.

[8] G. Piatetsky-Shapiro, and W.J. Frawley, Knowledge Discovery in Databases. *AAAI/MIT Press*, 1991.

[9] Q. Lu, and L. Getoor. Link-based classification. In *Proceedings of ICML-03*, 2003.

[10] Brin, and L. Page, The Anatomy of a Large Scale Hypertextual Web Search Engine,, Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998.

[11] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, Link analysis: Hubs and authorities on the world. Technical report: 47847, 2001.

[12] J. Hou and Y. Zhang, Effectively Finding Relevant Web Pages from Linkage Information, IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 4, 2003.

[13] Wang Jicheng, Huang Yuan, Wu Gangshan, Zhang Fuyan. Web mining: knowledge discovery on the Web. *Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference* on Volume 2, Page(s):137 - 141 vol.2 - 12-15 Oct. 1999

[14] Cooley, R.; Mobasher, B.; Srivastava, J.; Web mining: information and pattern discovery on the World Wide Web. *Tools with Artificial Intelligence,1997. Proceedings., Ninth IEEE* International Conference. Page(s):558 – 567 -3-8 Nov. 1997.

[15] Kleinberg, J.M., Authoritative sources in a hyperlinked environment. *In Proceedings of ACM-SIAM Symposium on Discrete Algorithms, 1998*, pages 668-677 – 1998.

[16] Data mining: Crossing the chasm, 1999. Invited talk at the 5th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining(KDD99).

[17] Charu C Aggarwal and Philip S Yu. On disk caching of web objects in proxy servers. In CIKM 97, pages 238{245, Las Vegas, Nevada, 1997.

[18] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proc. of the 20th VLDB Conference, pages 487{499, Santiago, Chile, 1994.

[19] Virgilio Almeida, Azer Bestavros, Mark Crovella, and Adriana de Oliveira. Characterizing reference locality in the www. Technical Report TR-96-11, Boston Uni- versity, 1996.

[20] Martin F Arlitt and Carey L Williamson. Internet web servers:

[21] Alex Buchner and Maurice D Mulvenna. Discovering internet marketing intelligence through online analytical web usage mining. SIGMOD Record, 27(4):54{61, 1998.

BIOGRAPHY

Mrs. P.Menaka, working as a Assistant professor, Department of Information Technology, Dr N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India

Ms.A.Prathimadevi, Pursuing M.Phil Research Scholar, Department of Computer Science, Dr N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India