

Pioneering the Classification in Deep Web Mining

Himani Sharma

PG Student

*Department of Computer Science and Engineering
Al-Falah University*

Aafrin Jamal

Assistant Professor

*Department of Computer Science and Engineering
Al-Falah University*

Abstract

The main purpose of this study is to discuss about web mining and its categories. Web mining is used to achieve information from web data. As now a days there are many users on the web which leads to amount of data generated. In this paper, web mining and its types will be discussed to clear the vision towards data on web how it is processed and managed. It comprises of three types: web content mining, web usage mining and web structure mining. This paper will also focus on analyzing the three types and enhancing their features and capabilities. Web mining technologies has played a vital role in discovering knowledge. Web mining has become more interesting now as it has grown fast in a short span of time in the field of research.

Keywords: Web Mining, Overview, Web Content Mining, Web Usage Mining, Web Structure Mining

I. INTRODUCTION

Now a days web is very important for everyone in today's life which is increasing rapidly fast bringing changes and new things for the users. Updation is necessary to fulfil the needs of the users daily. Web mining is used to extract the web information that is needed by the users so that the necessary details can be fetched and utilized. Automation is everywhere and in every field to avoid the human work in creation of anything. Web mining utilizes the automatic way of information extraction from the World Wide Web according to the preferences[1].

Web mining can be categorised into three which are the areas of research in the web mining field.

Firstly, there is Web Content Mining which is an application of data mining techniques to unstructured or semi-structured data, usually HTML-documents. Secondly, Web Structure Mining is a use of the hyperlink structure of the Web as an (additional) information source. Lastly, there is Web Usage Mining which is an analysis of user interactions with a Web server (e.g., click-stream analysis).

Web mining should be decomposed into these subtasks:[14]

- 1) Resource finding: the task of retrieving intended web documents.
- 2) Information selection and preprocessing: Automatically selecting and preprocessing specific information from retrieved web resources.
- 3) Generalization: Automatically discovers general patterns at individual websites as well as across multiple sites.
- 4) Analysis: Validation and/or interpretation of the mined patterns.

Web mining is the term of applying data mining techniques to automatically discover and extract useful information from the World Wide Web documents and services [2]. Although Web mining is very deep in the field of data mining but if compare both then data mining is not equal to web mining.

The unstructured feature of Web data triggers more complexity of Web mining. Web mining research is actually a converging area from several research communities, such as Database, Information Retrieval, Artificial Intelligence [3], and also psychology and statistics as well.

II. OVERVIEW OF WEB MINING

Web mining refers to the overall process of discovering potentially useful and previously unknown information or knowledge from the web data [4]. Web mining is used to capture relevant information, creating new knowledge out of the relevant data, personalization of the information, learning about Consumers or individual users and several others. Web mining uses data mining techniques to automatically discover and extract information from World Wide Web [5].

Many useful techniques like Information retrieval, Information extraction and machine learning have been used in the past in discovering new knowledge from the large amount of data which is available in the web.

These techniques have been compared with web mining [4]. Information retrieval works by indexing text and then selects useful information[6].

The three types of web mining in which Web content mining examines the content of the web as well as the result of the search. It can be further classified into Web page content mining and Search result mining. Web page Content mining is a traditional search of web page via content while search result mining is a further search of pages found from previous search [7]. Web Structure mining mines the structures like HTML or XML tags and gets information from the actual organization of the page. It uses interconnections between web pages to give weight to the page [7]. Web usage mining is the application of data

mining techniques to understand the web usage patterns. It mines data from logs, user profiles, user sessions, cookies, user queries, bookmarks, mouse click, scrolls etc. The three phases of web usage mining are preprocessing, pattern discovery, and pattern analysis [8].

III. WEB CONTENT MINING

Traditional technique of searching the web was via contents. Web Content mining is the extended work performed by search engines [9]. Web Content mining refers to the discovery of useful information from web content such as text, images videos etc. [11, 10]. Two approaches used in web content mining are Agent based approach and database approach [10, 9]. The three types of agents are Intelligent search agents, Information filtering/Categorizing agent, Personalized webagents [10]. Intelligent Search agents automatically searches for information according to a particular query using domain characteristics and user profiles. Information agents used number of techniques to filter data according to the predefined instructions. Personalized web agents learn user preferences and discovers documents related to those user profiles [10, 9].

To know the web content mining better and how it processes, let's take an example of any shopping site like jabong.com. There will be methods step by step firstly a web page of the shopping site will be displayed then it will be preprocessed where extraction of text will be done from HTML document. Now there will be vector creation where IR techniques will be applied with mining techniques and lastly information will be structured where data will be in the form of tables or any other structured formats.

This type of web mining helps the users in finding the appropriate information from the web as the information requested by them is gathered, categorized, organized and available on the world wide web. It's very easy and possible to determine the content for searching query. Information navigation is improved on the web which is provided by productive marketing. Web content mining is also useful as quality of information is very high which is helpful for the users. It is also helpful in maintaining information as in example of an online shopping site product's information is extracted, viewed and specified that is only through web mining.

IV. WEB USAGE MINING

Web Usage Mining is the discovery of meaningful patterns from data generated by client-server transactions on one or more Web localities.

Now a days there is a great increase in growth of business companies and their laws. Web is not only for transaction but also a place where in today's date millions of people are using web due to which large amount of data has been generated. Web usage mining helps in knowing about user like their behaviour on the web and how they use the web, means usage pattern which will give good results like dynamic content websites where personalization and customization are performed by finding many users with their access patterns and their navigational links. With the help of mined information from Web usage data helps in managing the site better which also give profit or an advantage to the network system in this discovery of process.

High level of web usage mining includes preprocessing, pattern discovery and pattern analysis which helps in many areas. Preprocessing or data preparation includes data cleaning where unnecessary or unwanted data or files is cleaned and filtration is done and the files are HTTP log files, then there comes user identification where user is identified through its ip address and then converted to domain name and further processing is done, after user identification, a session identification is presented through time intervals between the entries consecutively. Pattern discovery and pattern analysis are performed through techniques in various fields.

Web users can be interested in other techniques which can learn their information needs and preferences, which is user modeling possibly combined with content mining. On the other side, information providers can be interested in other techniques that can improve the effectiveness of the information on their websites by adapting the web sites design or by favouring the user's behaviour in satisfying the goals of the site.

V. WEB STRUCTURE MINING

Web structure mining is based on the hyperlink structure of the Web. If traditional processes and the assumed events are applied then that can result in negative. However, the appropriate handling of the links could lead to potential correlations, and then improve the predictive accuracy of the learned models [8].

Two algorithms that have been proposed to lead with those potential correlations: HITS and PageRank.

The main challenge for Web structure mining is to deal with the hyperlink structure 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 [8], which is situated 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.

As there is no single unit type structure or unifying structure there are different types of objects on the web with differences in the authoring style and content which is much greater than in traditional collections of text documents. The objects in the WWW

are web pages, and links are in-, out- and co-citation (two pages that are both linked to by the same page). Attributes include HTML tags, word appearances and anchor texts [8].

VI. CONCLUSION

Web mining is performing very well now a days in various fields of web. This paper shows the up-to-date things which is growing very fast through web mining types. Web mining is enhanced and paper shows that users are increased and they are using the web. Web usage mining and web content mining is discussed in brief that how the information is used and delivered.

REFERENCES

- [1] T. Sunil Kumar, Dr. K. Suvarchala, "A Study: Web Data Mining Challenges and Application for Information Extraction", IOSR Journal of Computer Engineering (IOSRJCE), Vol 7, Issue 3, Nov- Dec 2012, pp 24-29.
- [2] Oren Etzioni. The world wide Web: Quagmire or gold mine. Communications of the ACM, 39(11):65-68, 1996
- [3] R. Kosla, H. Blockeel. Web mining Research: A Survey
- [4] Kosla, R. and Blockeel, H. 2000. Web Mining Research: A Survey. SIG KDD Explorations. Vol. 2, 1-15.
- [5] Ajoudanian, S. and Jazi, M. D. 2009. Deep Web Content Mining. World Academy of Science, Engineering and Technology 49
- [6] Van. C. J. 1979. Information Retrieval. Butterworths.
- [7] Dunham, M. H. 2003. Data Mining Introductory and Advanced Topics. Pearson Education.
- [8] Srivastava, J., Cooley, R., Deshpande, M., Tan, P. N. 2000. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data.
- [9] Dunham, M. H. 2003. Data Mining Introductory and Advanced Topics. Pearson Education.
- [10] Inamdar, S. A. and Shinde, G. N. 2010. An Agent Based Intelligent Search Engine System for Web Mining. International Journal on Computer Science and Engineering, Vol. 02, No. 03.
- [11] Kosla, R. and Blockeel, H. 2000. Web Mining Research: A Survey. SIG KDD Explorations. Vol. 2, 1-15.
- [12] F. M. Facca, P. L. Lanzi, "Recent developments in Web Usage Mining Research" Artificial Intelligence and Robotics Laboratory
- [13] L. Getoor, Link Mining: A New Data Mining Challenge. SIGKDD Explorations, vol. 4, issue 2, 2003.
- [14] http://paginas.fe.up.pt/~ec/files_0506/slides/06_WebMining.pdf [Accessed on Feb. 6, 2013]