ISSN: 2321-9653

INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE AND ENGINEERING TECHNOLOGY (IJRASET)

Comparative Analysis of Web Mining Techniques: Survey

Kratika Srivastava[1]

M.Tech (CS&E)

School of Computing Science and Engineering, Galgotias University, U.P., India

***Abstract*—The cluster of technologies and design are known as web, which has now emerged as a fertile area for data mining research. The web mining research is being carried out across the globe. Many research communities such as database information retrieval, AI, natural language are working on it. In this survey paper, I have discussed various researches in the area of web mining and have suggested three web mining categories. I have also tried to point out the relationship between the web mining and the agent. I have represented some issues, the process, algorithms and the application of the work.**

***Keywords*— Web Mining, data mining, information retrieval, information extraction, Agent.**

1. INTRODUCTION

The World Wide Web collection of internet servers that support specially formatted documents. Documents support links to other documents, as well as audio, graphics, and video files. As we all know, the web is vast, diverse and dynamic and due to its vast and extra flexibility it has raised the scalability, multimedia usage and some temporal issues. Due to these factors web is facing information overloading [2]. Some issues are as follows:

a) Finding meaningful information:
When a user is searching for relevant information about a specific topic then irrelevant information is served more by a browser. This is a problem of low precision.
b) Difficulty in finding un-indexed information that may be useful.

Due to increase of information overloading it becomes difficult to map or index all information on the web that's the reason much of useful information gets dropped and remain un-indexed. There are so many techniques which are used to solve these issues such as database, information retrieval and natural language processing. Various web communities are also working on it.

This paper contains: Section 2, overview of web mining and classification of web mining. Section 3, 4, 5 will describe some research in respective categories. Section 6 related work and in section 7, future work.

2. OVERVIEW OF WEB MINING:

It is a mechanism or techniques to automatically extract and discover information from web documents [3]. Web mining can be studied in three different parts:

1) data extraction
2) Pre-processing.and Information selection
3) Generalization

ISSN: 2321-9653

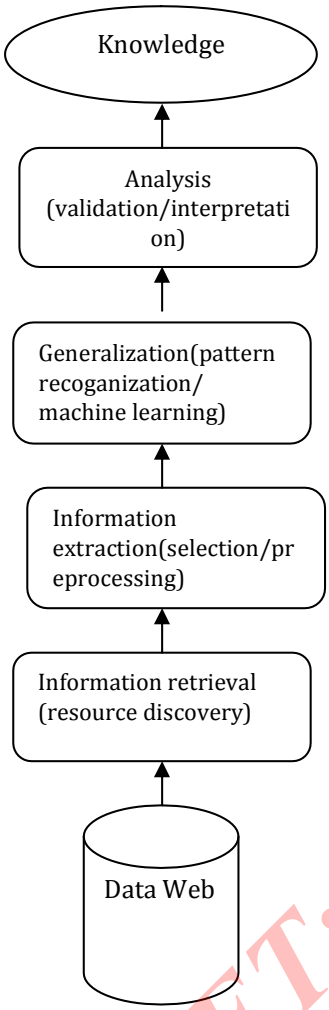

Figure 1: Process in web mining

❖ *Data extraction*: it is used to find the intended web documents or data which are actually required. Data can be available online or offline. Online data is available for research purpose. Or any information that is available via the use of www. Offline data can be hard copy of text, graph or any data available outside www.

❖ *Information selection and per-processing*: it is to select the useful information and then transform the original data. Information can go through a kind of pre-processing such as removing stop words, streaming etc.

❖ *Generalization*: discovery of patterns. E.g. - an email, email is classified as 'legitimate' or 'spam'.

❖ *Analysis*: it is to find validate data patterns produced by above steps. If desired results in these patterns are found then it is converted into knowledge. If not then they are further processed to obtain desired results.

Hence, I can say that, web mining is a process of discovering useful and unknown information or knowledge from the web data.

It includes standard process of database known as knowledge discovery in database (KDD) [2]. KDD is most used and most important in data mining or I can say that web mining is an extension of KDD.

2.1. *INFORMATION RETRIVAL:*

Information retrieval is an automatic retrieval process of documents or data or facts. An information retrieval system is a software programme that stores and manages information documents, often textual documents but possibly multimedia [4].

Its main aim is to search useful data collection. It also assists user in finding information which they are looking for. Information retrieval system gives relevant document and no irrelevant data can be found.

Information retrieval includes modelling, classification of information and then categorization further if required followed by data visualization and filtering [5].

2.2. INFORMATION EXTRACTION:

Information retrieval is concerned to identify relevant data from a document pile whereas information extraction produce structured data for processing [6]. We can say it the other way round that information extraction extract structured information from unstructured text [7]. It is based on web sites. Extraction can be automatically or semi automatically from the web. It can be taken as a pre-processing stage of web mining as it is performed after information retrieval and before web mining. Information extraction is very useful in natural language understanding, question-answering and summarization.

There are two types of information retrievals: unstructured and semi-structured. In unstructured, it performs linguistic pre-processing before data mining. Information retrieval from unstructured data can be classical or traditional

IE research. Information extraction depends on linguistic pre-processing which include semantic analysis, syntactic analysis and discourse analysis. Semi-structured data extraction is not based on linguistic restriction. This system uses Meta information such as an HTML tag. Because of the semi-structure nature of the Web, web content mining is different from text mining, while text mining focuses on unstructured texts. It requires creative applications of data mining or text mining techniques and also its own unique approaches. In the Web content mining area in the past few years there was a fast expansion of activities. However, the lack of structure of Web data and due to the heterogeneity automated discovery of targeted or unexpected knowledge information still presents many challenging research problems.

2.3. *MACHINE LEARNING APPLIED ON THE WEB*:

It allows computer to handle new situations via self-training, observation, analysis and experience [7]. It is an automatic method and it makes and improves predictions based on data. It is a great help to web mining and it gives advancement to new scenarios testing and pattern, adaptation and trend detection for improving decisions.

3. LITERATURE SURVEY

Many authors have been presented their survey and proposed their algorithm ever since for the web topology such as HITS [12], page rank [11] and improvements of hits adding content linking structure [13] and outlier filtering [14]. Some well-known examples are clever system [13] and Google [11].

In [6], the authors have discussed on the IE tools for semantic web and compared in many dimensions such as, the task domain, the techniques domain and the automation degree and also described why IE system fails in these dimensions. In [20], the author discussed how semantic web is differ from web 2.0 and also discussed advantages and disadvantages of both techniques. The author says that semantic web is differ but only needs some basic structure of the web to increase reliability and flexibility. In [21], first the authors discussed on semantic web and the proposed a model that gives the build process and complete description of functions of each module. The authors have given some outline to increase the scalability and to provide better result of semantic web [22].

In [16], Rajni Pamnani and Pramila Chawan discussed on web usage mining, approaches and applications of web usage mining. Govind Murari Upadhyay and Kanika Dhingra has been focused on web content mining and also discussed its application in current business environment. The author also explained how web content mining plays an integral role in decision making in the education, research and corporate environment.

Many authors have been published their paper in web mining and discussed on one or two web mining techniques but in this paper I gave the complete survey or briefly description of all three techniques of web mining such as: web content mining, web structure mining and web usage mining. This paper also contains its applications, advantages and disadvantages.

4. DESCRIPTION OF SELECTED TECHNIQUES

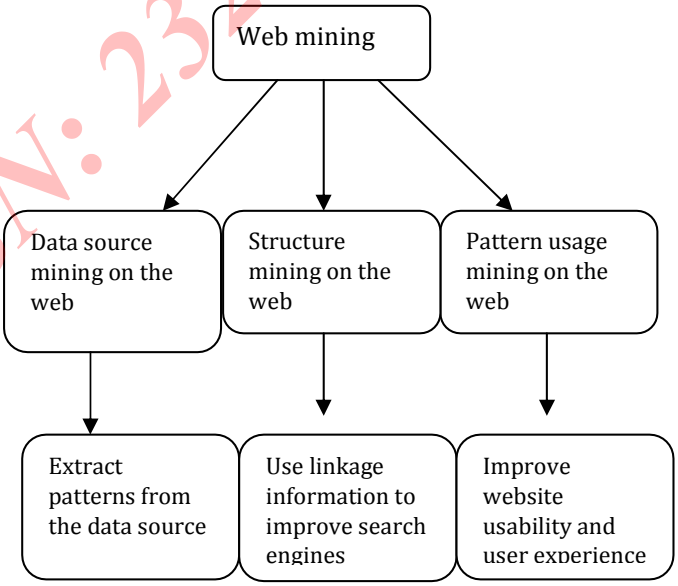

Figure 2: Taxonomy of Web Data Mining

3.1 *WEB CONTENT MINING:*

It is also known as text mining. It is scanning and mining of graphs of a web page, pictures and text to determine the exact content to the query. Previously the internet had services and data sources such as gopher, FTP and UseNet. Now every service and data sources are accessible via the web. A massive amount of information is available on the web, content mining provides results list to the relevance to the query [9]. The resultant pages or documents displayed by search engines are

according to the relevance i.e. from highest relevant information. These pages displayed have links so that users can fetch desired data from there. Therefore irrelevant information search is reduced. It removes the frustration and increase the navigation of data on the web. We could differentiate the research in web content mining from two different points: IR(information retrieval) and DB (database) views. IR from the view of web content mining is mainly to improve the information finding or filtering the data to the user. Whereas, web content mining in view of the database is that it tries to model data on the web and integrate them in more sophisticated queries. This is done by finding the schema of web documents, building a web warehouse. Semi-structured data has some structure but no rigid schema.

3.2 *WEB STRUCTURE MINING:*

It is a type of web mining and It is a tool among other mining tool which is used to discover the relationship between linked structures web or without linked structures. It is very useful to mine information such as the similarity and relationship between different web sites. According to database view I am looking for structures within web documents (intra-document structure). As we all have seen with in social network we find a variety of based on the outgoing and incoming links. This outgoing and incoming link analysis is very useful for social network model and to find the underlying links structure of the web itself [10].

3.3 *WEB USAGE MINING:*

Web usage mining is the third category in web mining. It provides path leading to access web pages. The information required is collected automatically into access logs via the web server. CGI scripts provide useful information such as reference logs, user subscription information and survey logs [15].web usage mining is very useful for companies and their intranet/internet applications and information processing. It is also beneficial in E-commerce and product oriented user services [16].

Web usage mining can be studied in two different commonly used approaches [17]. One is to draft the usage data on the web server into relational tables before any data mining technique is performed. Another approach uses the log data directly by using pre-processing technique. Data mining also uses association mining. Web usage data is can also be represented in graph form [19].

Web usage can be classified into two ways: i) knowledge about user profile ii) knowledge about user navigation pattern. An information provider is interested in techniques that could increase the effectiveness of the information on their web sites or I can say they are interested in navigation patterns. A tool such as system enhancement, personalization, site improvement, usage characterization and business intelligence improve the navigation patterns. Web usage data includes activities like browser logs, a web server access logs, and registration data. Proxy server data/logs, User profiles, user queries, cookies, registration data, transactions and mouse clicks etc. it also uses links, text and profiles like transaction records and business records etc. that are concluded by the user. The scope of web content mining from point of IR view is international while from a database point of view it's limited.

5. CONCLUSION

In this paper I have surveyed the research in the area of web mining. I have described three web mining categories and described some of the research associated with these mining categories. As a part of survey I have shown some representation issues, process, algorithm, applications etc. Some major issues on web information overloading have also been shown. Nowadays, web content mining is a very interesting research topic which could be a knowledge based or web warehouse. Finally another interesting topic which has emerged is that graph structures occurrence in web mining.

ACKNOWLEDGEMENT

I would like to thank my Guide Mr. Shiv Naresh Shivhare for his contribution towards the project. I would like to thank to my friend for providing me additional information regarding my project and helping me to complete this paper..

REFERENCES

[1] Faustina Johnson and Santosh Kumar Gupta, "Web Content Mining Techniques: A Survey", *nternational Journal of Computer Applications* 47(11):44-50, June 2012.

[2] U. Fayyad, S. Djorgovski, and N. Weir, "Automating the analysis and cataloging of sky surveys. In Advances in Knowledge Discovery and Data Mining", pages 471–493. AAAI Press, 1996.

[3] O. Etzioni, "The world wide web:", Quagmire or goldmine. Communications of the ACM, 39(11):65–68, 1996.

[4] Djoerd Hiemstra, "Information Retrieval Models", Published in: Goker, A., and Davies, J. Information Retrieval: Searching in the 21st Century. John Wiley and Sons, Ltd., ISBN-13: 978-0470027622, November 2009.

[5] R. Baeza-Yates and E. BerthierRibeiro-Neto, "Modern Information Retrieval" Addison-Wesley, Longman Publishing Company, 1999.

[6] Chia- Huichang, Mohammed Kayed, MohebRamzyGirgis and KhaledShaalan, "A Survey of Web Information Extraction System", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, TKDE-0475-1104.R3.

[7] "Information Extraction.pdf" by Broake Cowan, Oct 29, 2012.

[8] http://en.wikipedia.org/wiki/Semantic_Web accessed on 7th May 2014.

[9] http://en.wikipedia.org/wiki/Web_mining accessed on 7th May 2014.

[10] H. Kautz, B. Selman and M. Shah, "The hidden web", AI Magazine Volume 18 Number 2 (1997) (© AAAI).

[11] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine", Seventh International World Wide Web Conference, 1998.

[12] Prof. Ehud Gudes, "Graph and Web Mining -Motivation, Applications and Algorithms.pdf", Department of Computer Science Ben-Gurion University, Israel.

[13] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Mining the link structure of the world wide web", IEEE, 32(8):60–67, 1999.

[14] K. Bharat and M. R. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment" In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 104–111, 1998.

[15] Govind Murari Upadhyay and KanikaDhingra, "Web Content Mining: Its Techniques and Uses",International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, November 2013.

[16] "Web Usage Mining: A Research Area in Web Mining.pdf" by

RajniPamnani and PramilaChawan, Department of computer technology, VJTI University, Mumbai.

[17] J. Borges and M. Levene, "Mining association rules in hypertext databases", In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), 1998.

[18] RakeshAgrawal and RamakrishnanSrikan, "Mining sequential patterns", IBM Almaden Research Centre, CA 95120.

[19] A. B¨uchner, M. Baumgarten, S. Anand, M. Mulvenna, and J. Hughes, "Navigation pattern discovery from internet data", In Proceedings of the WEBKDD'99 Workshop on Web Usage Analysis and User Profiling, 1999.

[20] Parth R. Agarwal, "Semantic Web In Comparison to Web2.0", 2012 Third International Conference on Intelligent Systems Modelling and Simulation, DOI 10.1109, IEEE, 2012.

[21] WANG Yong-gui and JIA Zhen, "Research on Semantic Web Mining", 201O International Conference On Computer Design And Appliations (ICCDA 2010), IEEE, 2010.

[22] R. Guha, Rob McCool and Eric Miller, "Semantic Search", WWW2003, May 20-24, 2003, Budapest, Hungary. ACM 1.58113-680-3/03/0005.