An Analysis on Recent Reviews Pertaining to Web Mining and Its Application Domains

J.I. Christy Eunaicy[1], S. Suguna[2]
[1]Assistant Professor& Head, Department of Information Technology
Madurai SivakasiNadars Pioneer Meenakshi Women's College
Poovanthi, Tamil Nadu, India
[2]Assistant Professor& Research Supervisor, Department of Computer Science
Sri Meenakshi Government Arts College for Women
Madurai, Tamil Nadu, India

Abstract

*World Wide Web is a very popular and fertile area for interactive transformation of information. The Web Mining is mainly concerned with web content, web structure and web usagemining and extensive researches are being conducted in all these web mining categories. At this juncture, the pooling of reviews on web mining and their analyses are mandatory for finding new avenues and applications in researches. In this connection the present investigationwas devoted not only for the collection of reviewson mining of web usage, web structure and web content but also for the extensive analyses on the pooled reviews. Itwas found thatcontent mining in terms of scanning of text and mining of text, pictures plus graphs of a web page could determine the relevance of the content to the search query. It was also found thatcontent miningcould provide the list of results to search engines in the order of highest relevance to the keywords in the query.The research outcomes in structure mining revealed that the knowledge extracted from the web could be used to raise the performances for web information retrievals, question answering and Web based data warehousing. The result outcomes also revealed that the extraction of knowledge from server log files could be possible in web usage mining. On the basis of the pooled research findings in the present investigation, it could be concluded that the new avenues and applications in researches would be in increasing trends that would result in technically matured anduser friendly web mining applications***.**

Keywords: *Web Mining, Web Usage Mining, Pre-processing, Pattern analysis, Link Structure, Personalization.*

I. INTRODUCTION

Data mining is a broad field of recent researches and web mining is one of its major applications. While data mining is the process of collecting,searchingthrough,andanalysingalargeamountofdatainadatabase so as to discover patternsorrelationships in any organization, web mining is the process of using data mining techniques and algorithms to extract information directly from the web by extracting it from web documents and services, web content, hyperlinks and server logs [28]. It is reported that web content mining, web structure mining and web usage mining are the major web mining classifications [29]. It is also reported that extensive researches are being conducted worldwide in the applications of all these three classifications of web mining. At this juncture, it is essential to pool and document the reviews pertaining to web mining so as to make aware the recent researches to the researchers. It is also essential to analyse the documented reviews on web mining so that the application domain and research scopes can be explored and hence the present investigation.

II. REVIEW OF LITERATURE

The Web Mining is broadly classified as web content mining, web structure mining and web usage mining. Whiletheweb content mining is classified as web page content mining and search result mining, the web structure web mining is categorized into Link structure, internal structure and URL mining. At the same time, the web usage mining is categorized as web server logs and application server logs. In the present study, the relevant research reviews in each category have been collected from research journals and reports. The research results reported in reviews have been subsequently analysed. The application domains, scope of researches and expected research outcomes have been documented in this review paper.

Web Content Mining:

Web content mining is the process of miningof data from the content of web pages, extracting knowledge from the web contents and integration of useful data, information and knowledge from Web page substances. Content mining provides the results that are listed by search engines in the order of the highest relevance to the keywords in the query [2, 3]. It has the main tasks such as web page content mining and result page content mining. The review on web page content mining and result page content mining have been collected, pooled and analysed in the following sections:

Web Page Content Mining:

Web content mining is different from text mining through the structure of semi-structured web, while text mining is focused on unstructured text. Web content can be unstructured (eg text), semi-structured (HTML documents) or structured (dataextracted from databases in dynamic Web pages). The dynamic data cannot be classified, forming the so-called “hidden web”. In this context, Valarmathi and Purusothaman [4] reported that the content mining could also be related to text mining, because much of web contents were mostly text basedKao et al.,[5]dealt with a preliminary study on web page content mining and it was reported that collection of data was an important task, specifically, for web structure mining and content mining as that of data mining. It was also reported that a large number of target web pages were involved in the crawling process. Ozel [7] developed a genetic algorithm to select best features ofweb page classification. This algorithm could solve the problem to improve accuracy and run time performance of the classifiers. To determine whether a Web page belongs to a specific class (e.g. graduate student homepage, a course page, etc.) or not, a classifier needed “good” features extracted from the Web pages. As every component in a Web page such as HTML tags and terms could be taken as a feature, the dimension of the classification problem would become too high to be solved by well-known classifiers like decision trees and support vector machines. To decrease the feature space, the developed genetic algorithm should be used to determine the best features for a given set of Web pages. It was found that the accuracy improved upto 96% when features selected by this genetic algorithm were used and a kNN classifier was employed. PrabhjotKaur [6] proposed an approach to discover informative contents from a set of tabular documents of a web site by dynamically selecting the entropy threshold. It was noted thatthe system first partitioned a page into several content blocks according to HTML tag <TABLE> in a Web page. It was also observedthat the system was not applicable to general web pages which comprisedthe usage of tag<DIV>.

On the basis of the research outcomes on the collected reviews in connection with mining ofweb page content, itcan be concludedthat the collection of data plays a vital role in web content mining with the focus on unstructured, semi structured and structured contents and this predicted process shall provide required patterns to the end users. In the case of application domains, the web content mining can be applied in business sectors mainly for mining online news site and developing suggestion systems for distance learning. This application can help to establish better relationship with customer by providing exactly what they need. As far as the scope of research is concerned, the prediction analysis has to be improved in terms of usability and scalability

Web Structure Mining:

The knowledge extracted from the web can be used to raise the performances for web information retrievals, question answering and Web based data warehousing in web structure mining processed. As it is knownweb structure mining, one of three categories of web mining for data, is a tool used to identify the relationship between web pages linked by information or direct link connection. It offers information about how different pages are linked together to form this huge web. PreetibalaDeshmukhandVikramGarg[9]described the different algorithms which were used not only for the pageranking technique but also and to monitor the effective performance. Their contribution was to provide an efficient classification algorithm for the web pages which could help the user to search the result in efficiently as compared to the existing algorithm. Shivkumar and SeifedineKadry, Ali Kalakech and PardakheKeole, Malarvihi and Saraswathi[8,10,11,31]described that therewere some possible tasks of linkmining such as link based classification, cluster analysis, link type, link strength and link cardinality which would be applicable in web structure mining. It was noticed thatthe link based classification would be the most recent upgrade of a classic data mining task to linked Domains. The taskwould be beneficial to focus on the prediction of the category of a web page, based on words that would occur on the page, link between pages, anchor text, html tags and other possible attributes found on the web page. It was also noticed that the goal in cluster analysis would beuseful to find naturallyoccurring sub-classes. In link type, it was perceived that there would be a wide range of tasks concerningthe prediction of the existence of links, such as predicting the type of link between two entities, or predicting the purpose of a link. In linkstrength, it was per that the links would be associated with weights. At the same time, in link cardinality, it was noted down that the main task here was to predict thenumberof links between objects. Thereweresome uses of web structure mining like (a) Usage to rank the users queryand it would be segmented into groups, where similar objects could be grouped together,and dissimilar objects could begrouped into different groups. Different than the previous task, link-based cluster analysis should be unsupervised and could be used to discover hidden patterns from data and (b) usage for deciding what page would be added to the collection andpage categorization. This would be beneficial for finding related pages and also for finding duplicated web sites and similarity between them.It was reported that the structure mining could offer information about how different pages would be linked together to form huge web. It was also reported that different algorithms could be used for the page ranking technique and to monitor the effective performances. At the same time, the hidden patterns could also be discovered with the help of link-based cluster analysis.On the basis of the research outcomes on the collected reviews pertaining to mining of web structure, it can be concluded that page categorization can be added and related pages could be used to form huge web.

In the case of application, in the business world, structure mining can be quite useful in determining the connection between two or more business web sites. The connectivity shall be increased in the near future in terms of quality and quantity. As far as the research scope is concerned, it shall be with the structure of the hyperlinks with in the web itself.

***Web Usage Mining*:**

Web usage mining, is the application of data mining techniques and it is useful to find out interesting patterns from web usage data.

NamdevAnwatandVarshaPatil [12]mainly tried to extract useful and interesting patterns from usage data such as server logs, client browser logs, proxy server logs, cookies, user sessions, registration data, mouse clicks, user queries, bookmarks etc. and any other data as the results of user interactions. It is understood from the obtained result that trivial and useless knowledge could be distinguished so that further web modifications, system improvement and or web personalization could be executed.

AmitPratap Singh and Jain [13] explainedWeb Usage Mining could be that part of web mining, which could with the extraction of knowledge from server log files. It was stated that the server log files mainly consisted of the textual logs that were collected when users accessed web servers and might be represented in standard formats. It was also stated that the typical applications of web usage mining such as web personalization, adaptive websites and user modelling were to be analysed separately for finding out the scopes of research and extensive applications of the web usage mining.

Arvind K. Sharmaand Gupta [14] showed that the server log files were simple text files which could recordthe activity of the users on the server. Of course, these files resided on the server. If user visited many times on the Website, entry was created many times on the Server. The main source of raw data was the web access logs which were known as web server log files. The log files could be analysed over a time period and the time periodcould be specified on hourly, daily, weekly and monthly basis. The typical web server log files contained such type of information: IP address, request time, method (e.g. GET), URL of the requested files, HTTP version, return codes, the number of bytes transferred, the referrer's URL and user agents and the information found to be useful in web usage mining processes.

Maryam Jafari, FarzadSoleymaniSabzchi,ShahramJamali [15], examined the server log to remove the irrelevant and redundant items in the mining process. It was found that not only byfiltering out the useless data, but also by using log files the storage space could be reduced to facilitate the coming actions which could further reduce the size of web server log files. Singh and Arun[16],suggested that two kinds of records were unnecessary and they should be removed from the records of graphics, videos and the format information which had file name suffixes of GIF, JPEG, CSS and so on and the records with the failed HTTP status code.ChitraaAntony SelvadossThanamani[18]analysed entries with error and it was found that the status code could show the success or failure of a request. Entries with status code less than 200 and greater than 299 were found to be failure entries which had to be removed.SanjeevDhawanand SwatiGoel[19], dealt with rrequests that were performed by automated programs such as web robots plus spiders. The traffic that these programs generated could create false results and so such files had to be removed.Chaitra L Mugali et al., and RajniTripathi et al., [27, 30] suggested that the unwanted log records which were not useful for the further process should be removed from the web log file in Data Cleaning steps. It was reported that the records with ".jpg",".jpeg",".gif",".png","robout.txt","slurp","bot", "script","css”,".avi", ".js",extensionsshould beremoved from the input log file. Subsequently the processed files would be used for the further applications like Pattern Discovery and Pattern Analysis. On the basis of the results mentioned in reviews, it could be concluded that the unwanted entries should be removed from the web log file in the data cleaning process, so that the website would be ready for further application processes.

AkshayUpadhyay and BalramPurswan, SanjeevDhawan and Swati Goel, Sheetal A. Raiyani and, ShailendraJain and NirmalaHuidrom and NehaBagoria [17, 19, 20]analysed on user identification. The analyses revealed that the user identification was a complex job of web log pre-processing. And this task was essential to distinguish the users.The distinguishing task was due to the grouping of the users based on their visiting behaviour. As it is known, different techniques such as usage of IP address, referrer log and user agent could be used to identify the users. It was reported that the methods such as (i) Unique IP address could represent one user, (ii) If IP address would be same and agent log would be different, it would be considered as distinguish users.(iii) Construction of the browsing path by using the access log and referrer logs. It wasalso reported thatanother user in same IP address wasconsidered to identify the user, if there was mismatch in the browsing path.SugunaandSharmila[21] presented distinct user identification technique which was the enhancement of pre-processing steps of web log usage data in data mining. It was found that two pre-processing techniques could be used to combine within onepre-processing step time of user identification. It was also found out that the distinct user based on their attended session time would be required. As an algorithm for advanced pre-processing would be required, it was developed and it was found to be very efficient as compared to other identification techniques. Based on the obtained more precious and accurate results, one could easily personalizewebsites andimproves the design of web pages as usages of users on websites.

As far as pattern discovery was concernedSudheer Reddyet al., [24] showed the key principle involved in the method of pattern discovery. As a first step, extracting the sequential patterns on the original log had to be done.Subsequently, thesequential patterns had to be clustered and the web log had to be divided according to the clusters obtained. Finally, a distinct sub-log had to be created to collect the usersessions from the original sub-log which could not correspond to a cluster from the earlier step. In addition, the whole process had to be applied recursively for each sub-log.Etminani et al., [22] suggested, with the existing data of the log files many useful patterns had to be discovered either with user ids, sessiondetailsand time outs.

Chitraa,[23] suggested that the quality of a website could be evaluated by analysing user accesses of the website. To know the quality of a web site, user accesses had to be evaluated by web usage mining. The results of mining could be used to improve the website design and increase satisfaction which could help in various applications. It was reported that log files were found to be the best source to know user behaviour. But, the raw log files could contain details like image access and failed entries, which would affect the accuracy of pattern discovery and analysis. So, the pre-processing stage was noted to be an important work in mining to make efficient pattern analysis. It was also reported that the user’s session details had to be known to get accurate mining results. Sisodia[25] described the web usage pattern analysis as the process of identifying browsing patterns by analysing the user's navigational behaviour. The web server log files which stored the information about the visitors of web sites were used as input for the web usage pattern analysis process. In this connection, these log files were pre-processed and converted into requiredformats. Subsequently,the web usage mining techniques were appliedon these web logs.KobraEtminani et al., [22] carried out the advantages of pattern analysis and proposed novel methods. In these methods, suitable dissimilarity function for the clustering step was required and it was noted that complex mathematical relations were not required. It was deduced that other similar methods were limited in terms of the length of the extracted patterns.In the research article, it was noted that it would be possible to extract patterns of any length.It was also reported that different patterns could be extracted depending on theoccurrence of that page group in a cluster. It was noticed that even pagegroups that were less accessed could be extracted.It was noticed that these kinds of results would be useful for web site owners.Aarti Parekh et al., [26]proposed a system which could discoverthe useful pattern from web server log file. In the case of webtransactions, association rules were found to have relationships amongpage views based on the navigation patterns of users. The implementation of a priori algorithm on the web log files would give frequently accessed webpages and unique users.

On the basis of research outcomes related to mining of web usage it is concluded that web usage mining plays an important role in realizing enhancing the usability of the website design, the improvement of customers' relations and improving the requirement of system performance and so on.The applications of web usage mining are like improving website design, improving performance of system, pre-fetching and caching. Semantic web is the future scope in web usage mining.

III CONCLUSION:

This paper discussed a survey of framework on web mining. In the case of application, in the business world, structure mining can be quite useful in determining the connection between two or more business web sites. The connectivity shall be increased in the near future in terms of quality and quantity and also the web usage mining plays an important role in realizing enhancing the usability of the website design, the improvement of customers' relations and improving the requirement of system performance and so on.The collection of data plays a vital role in web content mining with the focus on unstructured, semi structured and structured contents and this predicted process shall provide required patterns to the end users.

REFERENCES

[1] .Srividya M, Anandhi D, Irfan Ahmed M.S., ”Web Mining and Its Categories – A Survey”. International Journal Of Engineering And Computer Science ISSN: 2319-7242, Volume 2 Issue 4 April, Page No. 1338-1345, 2013.

[2]Rajdeepa B, Sumathi P., “An Analysis of Web Mining and its types besides Comparison of Link Mining Algorithms in addition to its specifications”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3, Issue 1, , ISSN: 2278 – 1323, January 2014.

[3] Neha Sharma et al., ‘’ A Hand to Hand Taxonomical Survey on Web Mining”, International Journal of Computer Applications (0975 – 8887) Volume 60– No.3, December 2012

[4] Valarmathy et al.,” A Survey on Web Content Mining Techniques and Tools” IJISET, Vol. 1 Issue 6, , ISSN 2348 – 7968, August 2014.

[5] Kao et al., “Wisdom Web intrapage informative structure mining based on document object model” in IEEE Trans KDD, 2005.

[6]PrabhjotKaur."Web Content Classification: A Survey". International Journal of Computer Trends and Technology (IJCTT) V10(2):97-101, ISSN:2231-2803, 2014.

[7] Ozel, S.A, A genetic algorithm based optimal feature selection for Web page classification, IEEE, pg.no 282-286, ISBN:978-1-61284-919-5

[8]Shivakumar et al., “Survey On Web Structure Mining” ARPN Journal of Engineering and Applied Sciences, Vol. 9, No. 10, , Issn 1819-6608, October 2014.

[9] PreetibalaDeshmukh et al., “A Survey Paper of Structure Mining Technique Using Clustering and Ranking Algorithm”, International Journal of Computer Applications (0975 – 8887)Volume 119 –No.13, June 2015

[10] SeifedineKadry , Ali Kalakech ,” On the Improvement of Weighted Page Content Rank”,Journal of Advances in Computer Networks, Vol. 1, No. 2, June 2013

[11]Pardakhe et al., “Analysis of Various Web Page Ranking Algorithms in Web Structure Mining”International Journal of Advanced Research in Computer and Communication Engineering Vol.2, Issue 12, December 2013

[12] NamdevAnwat et al.,”Survey paper on web usage mining for web personalization”, International journal of Innovative Research and Development, ISSN 2278–0211, Vol 3, Issue 7, 2014

[13] AmitPratap Singh et al.,” A Survey on Different Phases of Web Usage Mining for Anomaly User Behavior Investigation, International Journal of Emerging Trends & Technology in Computer Science , Volume 3, Issue 3, ISSN 2278-6856, May – June 2014.

[14] Arvind K. Sharma et al., “Analysis Of Web Server Log Files To Increase The Effectiveness Of The Website Using Web Mining Tool”, International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 4, Issue 1, pp1-8, 2013.

[15]Maryam Jafari1 et al., “Extracting Users' Navigational Behavior from Web Log Data: a Survey”, Journal of Computer Sciences and Applications, Vol. 1, No. 3, 39-45, 2013.

[16] Singh et al., “Web usage Mining: Discovery of Mined Data Patterns and their Applications”, International Journal of Computer Science and Management Research, Vol 2 Issue 5, ISSN 2278-733x, 2013.

[17]AkshayUpadhyay et al., Web Usage Mining has Pattern Discovery”, International Journal Of Scientific and Research Publications, Volume 3, Issue 2, February 2013.

[18] Chtiraa et al., “A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing”, International Journal of Computer Applications (09275-887), Vol 34 – No.9, November 2011.

[19] SanjeevDhawan et al., “Web Usage Mining: Finding Usage Patterns from Web logs”, American International Journal of Research in Science, Technology, Engineering & Mathematics, ISSN 2328 – 3491, pp 203-207, 2013.

[20] Sheetal A. Raiyani et al., “Efficient Preprocessing technique using Web log mining, International Journal of Advancements in Research & Technology”, 1(6) ISSN 2278-7763, 2012.

[21] Suguna et al., “User Interest Level Based Preprocessing Algorithms Using Web Usage Mining**”,** ISSN : 0975-3397 Vol. 5 No. 09 Sep 2013.

[22] KobraEtminani, Mohammad et al., “Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method,” in Proc. IFSA-EUSFLAT 2009.

[23] Chitraa et al., “A Survey on Preprocessing Methods for Web Usage Data “, (IJCSIS) International Journal of Computer Science and Information secretary Security, Vol. 7, No. 3, 2010.

[24] Sudheer Reddy et al., “An Effective Methodology for Pattern Discovery in Web Usage Mining”, International Journal of Computer Science and Information Technologies, Vol. 3 (2) 3664-3667, ISSN: 0975 – 9646, 2012.

[25**]**Sisodia, “Web usage pattern analysis through web logs: A review”, Computer Science and Software Engineering (JCSSE), 2012 International Joint Conference on May 30 2012-June 1 2012, pg:49 – 53 ISBN:978-1-4673-1920-1, 2012.

[26]Aarti M. Parekh et al., “Web usage Mining:Frequent Pattern Generation using Association Rule Mining and Clustering“, International Journal of Engineering Research & Technology (IJERT)ISSN: 2278-0181 Vol. 4 Issue 04, April-2015

[27] Chaitra L Mugali et al., “Pre-Processing and Analysis of Web Server Logs”, International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163Issue 8, Volume 2, 2015.

[28] RajniPamnani, PramilaChawan 1 Qingtian Han, XiaoyanGao, “Web Usage Mining: A Research Area In Web Mining”**.**

[29]Wenguo Wu, “Study On Web Mining Algorithm Based On Usage Mining”, Computer- Aided Industrial Design And

Conceptual Design, 2008. CAID/CD. 9th International Conference On 22-25, 2008.

[30] RajniTripathi, Munesh Chandra Trivedi, ShraddhaTripathi, "Web usage mining: A fact finding approach in web mining", International Journal of Computer Trends and Technology(IJCTT), vol12 no 12, June 2014

[31] Malarvizhi, Saraswathi "Web Content Mining tools and algorithms – A Comprehensive study", ", International Journal of Computer Trends and Technology(IJCTT), vol 4 no 8, 2013