ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 15, March 2017

Identification of User Interests using Navigational Patterns

Lidiya Nixon[1]
Computer Science Department
M A College of Engineering
India
lidiya.nixon@gmail.com

Linda Sara Mathew[2]
Computer Science Department
M A College of Engineering
India
lindasaramathew@gmail.com

***Abstract*—** The usage of internet is increasing day by day. Internet availability posed to be a problem during the earlier days. Today in this modern world where internet is always available at fingertips the main problem faced by users is that of getting interested data at the right time. Here a novel method is proposed to solve this issue. The method involves prioritization of web links followed by a navigational pattern prediction based on the search history of the user. The prediction is done using a TF-IDF score calculation. The cosine similarity matrix is then used to decide the preferred links for the selected user. If a user is able to get the desired links without much browsing this can increase the searching efficiency to a great extent.

Keywords— ***Data Mining; Web Mining; Genetic Algorithm; Navigational Patterns; TF-IDF; Cosine Similarity***

I. INTRODUCTION

World Wide Web has brought revolutionary changes in the popularity of internet. It has grown into a huge and global information space. The volume of information present on the web is distributed in nature and growing at an exponential rate. To get the desired information without wandering through the pages of website has become an irksome job. Different types of methods are required to organize and manage the information so that it can be used efficiently for business purpose. There exists a need of web mining technique in order to explore such a gigantic information base. Web mining is the application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc. Two different approaches were taken in initially defining web mining. First was a "process-centric view," which defined web mining as a sequence of tasks. Second was a "data-centric view," which defined web mining in terms of the types of web data that was being used in the mining process. The second definition has become more acceptable, as is evident from the approach adopted in most recent papers. So broadly speaking web mining is the application of data mining techniques to extract knowledge from web data, i.e. web content, web structure, and web usage data.

Based on the type of data present in web documents, web mining is divided into three classes: web content mining, web structure mining and web usage mining. Web content mining searches the information from structured, semi structured or unstructured content of the web. There are a number of links present on the web pages which connects and organizes the web structure mining for retrieval of information. Web usage mining discovers the usage pattern of visitor by mining the log files. It works by pre-processing the initial log data which removes the redundancy among data and then detecting the patterns and then performing an analysis on these patterns in order to find out user behaviour.

Several optimization techniques have been used for find the most useful pages of web site by using web usage and web content mining. The proposed approach uses natural optimization technique called genetic algorithm to explore the search space by using both content and usage mining. The inspiration behind genetic algorithm is the process of natural selection and genetic dynamics. Genetic algorithm has its roots in the Darwin's theory of survival of the fittest. So genetic algorithm is a search algorithm based upon the process of natural selection and population genetics. The proposed approach aims to use genetic algorithm on the data collected by integrating web usage mining and web content mining in order to find the pages of web site which are of utmost importance to user.

The proposed system consists of web log mining and online navigational pattern prediction. The usage data selected to work on is the data logged by a web server in a file called

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 15, March 2017

access log file. Usage data can also be collected from visitors' computers by using adapted browsers or by using techniques such as cookies; which might raise privacy concerns.

The web links prioritization and pattern prediction system starts by creating page views and ends up by generating sessions. Amongst the existing approaches for sessions' identification are the time-based heuristics. The idea of these time based heuristics is the use of a duration threshold to decide whether a session has ended or not. In order to mine for navigational patterns it is mandatory to know what visitors have looked at each time they have visited the website. Each time a visitor comes to the website is considered a session. Identifying users' sessions from the web log is not easy as it may seem. Logs may span long period of time during which visitors may come to the website more than once. Therefore, sessions' identification becomes the task of dividing the sequence of all page requests made by the same user during that period into subsequences called sessions. Many approaches have been used by researchers for sessions' identification. The most popular session identification techniques use a time gap between requests. It has been mentioned that that many commercial products use 30 min as default time-out threshold. However, many thresholds can be found in the literature. These thresholds vary from 10 minuites to 2 hours. It has been also mentioned that the most widely used time gap is 25.5 min established based on empirical data.

The combination of both prioritization and online navigational prediction is supposed to increase the efficiency of search engine to a great extent. [9] discussed about a system,the effective incentive scheme is proposed to stimulate the forwarding cooperation of nodes in VANETs. In a coalitional game model, every relevant node cooperates in forwarding messages as required by the routing protocol. This scheme is extended with constrained storage space. A lightweight approach is also proposed to stimulate the cooperation.

II. RELATED STUDIES

Web usage mining is the most crucial field of web mining. A lot of research has been done in this area which shows the importance of web usage mining to search engines. Speed and precision acts as most desirable characteristics of search engines.

The efficiency of search engine can be improved through web usage mining by using MASEL (matrix analysis on search engine log) algorithm proposed in [3]. The relationship among user, query and resource acts as central idea for this algorithm. MASEL considered a resource to be good if it is accessed by many good users.

Pattern discovery is performed in order to draw useful patterns from preprocessed data [4]. A system called Web Sift is designed to perform usage mining. It utilizes data from web server log in order to perform mining task. This data suffers from real world challenges.

A system named Predicting User Navigation Patterns Using Clustering and Classification later evolved [5]. It contained preprocessing the web log data and then identification of potential users. Then the identified users were clustered and a prediction model was created. Then based on user requests submitted to the prediction engine and future request prediction based on given request was done.

An attempt was also done to prioritize the ordering of URL queue in focused crawler [6]. For a crawler, it is not a simple task to download the domain specific web pages. This unfocused approach often shows undesired results. Therefore, several new ideas have been proposed, and crawling is a key technique, which is able to crawl particular topical portions of the World Wide Web quickly without having to explore all web pages. Focused crawling is a technique, which is able to crawl particular topics quickly and efficiently without exploring all webpages. The proposed approach does not only use keywords for the crawl, but also rely on high-level background knowledge with concepts and relations, which are compared with the texts of the searched page.

III. PROPOSED METHODOLOGY

The basic architecture of proposed system is as shown in Figure 1. The system mainly contains two modules. They are Web Links Prioritization and Online Navigational Pattern Prediction.

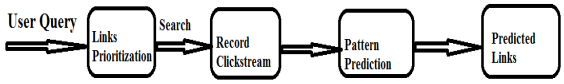

Figure 1 Basic Architecture

The proposed system introduces the prioritization of web links based on a certain set of factors which can make ranking a link for priority more efficient. Once the prioritized links are returned to user based on the query then comes the process of searching where user uses these links. The click stream of user is recorded. This order is compared with the server log to identify similar patterns of search. The server log is divided

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 15, March 2017

into sessions for this purpose. Then similar sessions are identified. Once similar sessions are identified then comes the process of pattern prediction. From the similar sessions identified the session which has the greatest similarity with the current user session is identified. Then the links in this session will be suggested to the current user.

For the purpose of web links prioritization genetic algorithm is used. The Genetic Algorithm (GA) is a natural optimization and adaptive heuristic search technique whose basic idea depends upon process of natural evolution. The mechanism of evolution is parallel in nature and has been used for solving several computational problems. GA is used for solving general purpose optimization problems. The Genetic Algorithm (GA) is a natural optimization and adaptive heuristic search technique whose basic idea depends upon process of natural evolution. The mechanism of evolution is parallel in nature and has been used for solving several computational problems. GA is used for solving general purpose optimization problems. In computational problem genetic algorithm begins by selecting initial population in the form of chromosome and then applying fitness function which minimizes the cost on selected chromosome. Then two parent chromosomes having greater fitness are selected. Crossover and mutation are performed on selected parents. The process is repeated until best solution among current population is retrieved. After selection crossover is performed between two parent string and it results into offspring string. Mutation is another operator which is applied after crossover in order to change genetic material between parents and forms offspring. Then on the basis of Darwinism the offspring which survives most is chosen to be fittest.

Various parameters are required for calculating the fitness of a solution as presented below.

a) Access frequency: Access frequency measures number of times a particular page is visited by user irrespective of user id in web usage mining, the usefulness of any particular page can be measured by calculating the access frequency. More the access frequency more could be its usefulness.

b) Number of unique visitors: This factor shows the importance of any web page on the basis of unique visitors who visited the page. This means that a URL can have more popularity among users if it is visited more by more number of distinct visitors.

c) Time Duration: The amount of time spent on a page shows the relevance of page for the user. If a user spent more amount of time on a particular page then that page is considered to be useful for the user.

d) Number of bytes received: The quantity of data downloaded by user from the web page shows that page has content which is relevant for user. The entries for number of bytes received by user are present in web log server entry. From this entry it can deduced if a page is important or not.

The fitness function is evaluated based on these parametrs. Then a binary tournament selection will be done to select the parents. Then the process of crossover is done on selected parents. And finally the mutation operation is done to introduce variants.

Once the set of prioritized web links are returned to user based on a query then the user uses these links. The user clicks these links in a particular order. This click stream is recorded for the purpose of pattern prediction. The order of links used by the user is obtained from the recorded click stream and is very useful for pattern prediction. This shows the pattern of current online session. This is compared with the server log.

For the purpose of pattern matching initially the server log file is divided into different sessions. For this initially sessions are divided into sub sessions based on a time lag say 30 minutes. Once sessions are formed with this time lag then adjacent sessions are compared for existence of similar pages. If more than a specified number of similar pages are found then such sub sessions are combined into a single session. This can avoid pages in same session being grouped into different sessions due to problems like network errors, time out errors etc. Testing whether a list of sessions share a pattern or not needs to be performed in an efficient manner.

The pattern prediction system consists of the adoption of the well-known technique used in information retrieval systems, namely TF-IDF combined with the cosine similarity measure to find the closest sessions to a current online session.

Here online pattern detection is viewed as the problem of finding the most relevant documents to a query made of set of keywords from a repository of text documents based on the following observations. A text document is made of a collection of some of the words available in the vocabulary of a specific language. Similarly, a session is made of a collection of the references of some of the pages of a website. Text documents may contain repeated terms. Also, a session may contain one page or more many times. This can happen simply when a user reload a page or comes back to one of the pages viewed earlier. Therefore, based on the aforementioned observations, sessions are treated as text documents and they benefit from the well-developed techniques in the informational retrieval domain to increase results accuracy.

When a visitor is browsing the website, only the last few requested pages are kept track of. The concept of sliding window is used. Each time the visitor requests a new page the windows is slide by one. Consequently, the newly requested page will be added and the oldest page in the window will be dropped. Assume that the size of the sliding window is w. Before any prediction can be made about the navigation pattern of the current visitor, the system should wait till the visitor requests at least w pages. Once it is the case, a query

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 15, March 2017

vector is created. The size of the vector is the number of pages in the website. The values in the query vector are the TF-IDF values of each requested page existing in the current sliding window. If a page does not exist in the current sliding window then its TF-IDF value is zero. The next step is computing the cosine between the query vector representing the online session of the current visitor and a selected sub-list of vectors representing relevant sessions to the current online session. Relevant sub-list of vectors can be obtained by using an inverted index. Therefore, instead of computing the cosine between the current online session and all sessions, it is done only for a subset of the sessions. A TF-IDF value of a specific term in a document d is calculated as the following:

TF-IDF(d,t)=[log(1+n(d,t)/n(d)]/n(t)

where n(d, t) is the number of occurrences of term t in document d; n(d) is the number of terms in document d. n(t) is the number of documents containing term t. The use of TF-IDF and cosine similarity method improves the efficiency of pattern prediction to a great extent.

This method can increase the efficiency of search engine to a great extent since the search results are made better by both prioritization and prediction.

IV.EXPERIMENTAL RESULTS

The system was tested by submitting different user queries. The efficiency of system is supposed to increase as the search continues. This is because more searching introduce more search patterns which makes comparison more efficient.

V.CONCLUSION

The web links prioritization and pattern prediction system introduces a hybrid system which combines prioritization and pattern prediction. The process of prioritization is achieved using genetic algorithm. This provides high quality web pages as the initial result of query. Each of these web links will be important in some specific manner. Then as the user continues his search the sessions similar to current online session are identified from server log. From these identified sessions the pages from most similar session is suggested to the user. This can improve the search engine performance to a great extent. In the modern world where the efficiency of search engine is most important the approach introduced here can greatly enhance the performance thereby making web surfing more efficient and accurate.

References

[1] Kamika Chaudhary, and Santhosh Kumar Gupta. "Prioritizing web links based on web usage and content data." 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT).

[2] Abdelghani Guerbas, Omar Addam and Omar ZaarourWangming. " Effective web log mining and online navigational pattern prediction." Computer Science and Software Engineering,2014,Knowledge Base Systems.

[3] D. Zhang, and Y. Dong, " A novel web usage mining approach for search engines," Computer Networks, vol 39(3) ,pp 303-310, 2002.

[4] Srivastava, R. Cooley, M. Deshpande and P.N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data", ACM SIGKDD Explorations Newsletter, 1(2), pp.12-23, 2000

[5] Samir S. Shaikh and Pravin B. Landage, "User Navigation Pattern Prediction using Longest Common Subsequence", International Conference on Recent Trends in engineering & Technology - 2013(ICRTET'2013).

[6] D. Koundal,"Prioritizing the ordering of URL queue in focused crawler", Journal of AI and Data Mining,Vol 2,No.1,2014,25-31.

[7] F. Picarougne, N. Monmarche, A. Oliver and G. Venturini, "GeniMiner: Web Mining with a Genetic-Based Algorithm," ICWI, pp. 263-270,2002.

[8] W. Fan, M. Gordon and P. Pathak, "Genetic programming-based discovery of ranking functions for effective web search," Journal of Management Information Systems, vol 21(4), pp 37-56, 2005.

[9] Christo Ananth, M.Muthamil Jothi, A.Nancy, V.Manjula, R.Muthu Veni, S.Kavya, "Efficient message forwarding in MANETs", International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE), Volume 1,Issue 1, August 2015,pp:6-9

[10] S. P. Nina, M. Rahman, K. l. Bhuiyan and K. Ahmed, " Pattern discovery of web usage mining," In Computer Technology and Development, ICCTD 09 International Conference on vol. 1, pp. 499-503 IEEE 2009

[11] S. K. Pal, V. Talwar, and P. Mitra, "Web mining in soft computing framework: Relevance, state of the art and future directions," Neural Networks, IEEE Transactions ,vol 13(5), pp.1163-1177, 2002.

[12] C. C. Lin, "Optimal Web site reorganization considering information overload and search depth," European Journal of Operational Research 173(3), pp.839-848, 2006.

[13] M. Mitchell, " An Introduction to Genetic Algorithms," MIT Press. Chapter 1-6. pp. 1-203, 1998

[14] A. K. Mishra, M. K. Mishra, V. Chaturvedi, S. K. Gupta and J. Singh, "Web usage mining using self organized maps" International Journal of Advanced Research in Computer Scence and Software Engineering, vol3(6), pp. 532-539, 2013

[15] M. Agosti, G.M.D. Nunzio, Web log mining: a study of user sessions, in:Proceedings of the 10th DELOS Thematic Workshop on Personalized, June2007

[16] Ester M., Gro M. and Kriegel H.-P.: 2001, Focused Web crawling: A generic framework for specifying the user interest and for adaptive crawling strategies, Technical report, Institute for Computer Science, University of Munich

[17] Deepika Koundal, Mukesh Kumar, Renu Vig, "Prioritizing the URLs in Ontology based Crawler" published and presented at International Conference of IEEE- AICC '2009 at Thapar University, Patiala.

[18] Page, L., S. Brin, R. Motwani, T. Winograd. "The PageRank Citation Ranking: Bringing Order to the Web", Stanford Digital Library Technologies Project.

[19] Debashis Hati, Amritesh Kumar, Lizashree Mishra, 2010. Unvisited URL Relevancy Calculation in Focused Crawling Based on Naïve Bayesian Classification, International Journal of Computer Applications, Volume 3- No.9.

[20] V. Sujatha, Punithavalli, "Improved user navigation pattern prediction technique from web log data", International Conference on

Communication Technology and System Design, 2011, Elsevier publication, Procedia Engineering 30 (2012) pp. 92 – 99

[21] Yue-Shi Lee, Show-Jane Yen, "Incremental and interactive mining of web traversal patterns", 2008, Elsevier publication, Information Sciences 178 (2008) pp.287–306.

[22] Neetu Anand, Saba Hilal, "Identifying the User Access Pattern in Web Log Data", International Journal of Computer Science and Information Technologies, Vol. 3 (2) , 2012, pp.3536-3539

[23] Yan Wang, "Web Mining and Knowledge Discovery of Usage Patterns" Google Documents, 2000.

[24] Kumar, P.R. and Singh, A.K., "Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval", American Journal of Applied Sciences, 2010, Vol. 7, No.6, Pp. 840-845.

[25] Haibin Liu, Vlado Kes˘elj, "Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests", Data & Knowledge Engineering 61 2007, Elsevier publication, pp. 304–330.

AUTHOR PROFILE

Lidiya Nixon is currently pursuing M.Tech in Computer Science and Engineering at Mar Athanasius College of Engineering, Kothamangalam.

Linda Sara Mathew is currently working as Assistant Professor at Mar Athanasius College of Engineering,Kothamangalam