IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

A Survey of Web Log for Effective Information Retrieval Using Data Mining Techniques

T. Nandhini[1], Dr. B. Kalpana[2]

Research Scholar, Department of Computer Science, Avinashilingam Institute of Home Science and Higher Education for Women, Coimbatore[1]

Professor, Department of Computer Science, Avinashilingam Institute of Home Science and Higher Education for Women, Coimbatore[2]

Abstract: With the improving communication technology, the people of the world need more information, to gather their information through the web. With the increase number of web user, the web search engines have to provide a better communication mechanism to those people. During their surfing their session may have the timed out with the fixed threshold mechanism. And with the help of the web log it can obtain it back. Hence it is the great problem for the web user. The searching data may arise with the help of the frequently searched items. With the help of these web logs and statistical language modeling, the user can get the relevant data. This paper provides a various technique of web log mechanism and helps the web user to surf more efficiently.

Keywords: WWW, Web Usage Mining, Web Log, Information Retrieval and Association Rules.

I. INTRODUCTION

World Wide Web (WWW) is growing very leading every day in the measure of websites and also the populace of users. It serves a large quantity of global information services for news, client information, education, finance management, and also many other services.

Web mining is a branch of data mining concentrating on the World Wide Web as the primary data source. The process of extract useful information directly from the internet and to seem for knowledge patterns in web data by aggregation and analyzing information in order to achieve insight into trends, the industry and users in general. Web mining is used for web personalization, system improvement and site modification.

In general, web mining is categorized into three types shown in Fig.1. Such as,

- Web content mining,
- Web structure mining and
- Web usage mining.

A. Web content mining
Web content mining, additionally referred to as text mining, is usually the second step in internet data processing.

Content mining is the manner of extracting beneficial records from the contents of internet documents. Content data is the gathering of information in a web page such as, textual content, photographs, audio, video, or structured information which includes lists and tables of an internet page to work out the connection of the content to the search query.

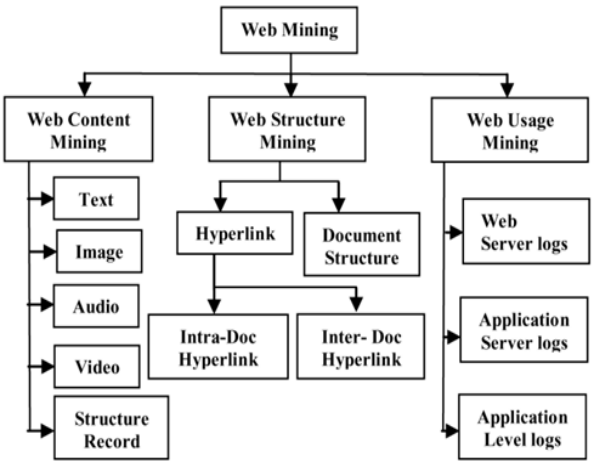

Fig.1 Web Mining

B. Web structure mining
Web structure mining is the tool used to determine link between web sites connected by data or direct link connection. There are two groups of web content mining strategy they are, the first strategy is directly mining the content and the second strategy is to enhance on the content search of other tools like search engine.

C. Web usage mining
Web Usage Mining is the procedure of extracting knowledge from Web users' access data by exploiting Data Mining technologies. It may be used for various functions such as personalization, system improvement and website modification. Web usage mining consists of three phases such as, preprocessing, pattern discovery, and pattern analysis. The basic rationale of website is to deliver functional information to its users competently and

appropriate; at the similar time websites are rival to obtain their own split of visitors. Websites are determined to get better themselves by contribution modified inside and services that apparently will match best of the users' flavor or needs. It is fine know that users' online connections with the website are evidence in server web log files that supply round as a priceless pool of in sequence. By be appropriate the data mining techniques on web log file, we gain good impending about the users' behaviors; thereby we can make to order the filling and services on the website to superior suit the users. We can scrutinize the web log files for various characteristic of website augmentation Web mining is an imperative research restraint of data mining and illustration huge concentration from academia and software manufacturing. It is a significance part of the online Knowledge Discovery development where data mining method are harvesting knowledge over the data together from World Wide Web [1]. One of the most imperative fields of the Informatics, and one of a smaller amount studied, is the analysis of the in sequence usage of web sites. During the examination of web log data we can scrutinize the activities of a web surfer in a web site, remove navigational outline about their favorites pages, recitation the used paths in organize to access to relevant in sequence and inspection the consistency of our web design and planning. Weblog analysis found statistical differentiation between the originated sessions from poles apart point of access. Thus, sessions imminent from search engines are longer in number of clicks than the assembly from the web site root [2]. Comparable result was originate with note that the useless time by a user that come from a search engine, a linkage or the root. Search engine's users offer more time to our web site that user imminent from other summit of access.

One imperative subarea is Web usage mining, wherein one attempt to determine patterns of Web usage from Web record data. The Web log data are typically noisy and predominantly confusing. There vestiges potential for discovering useful structure in the communication between a Web site and its users. Such data can be premeditated to produce supposition about Web site design, test example of Web sites or their adjustment and experiment hypotheses about the possessions of different design variables on Web-user behavior. Web logs confirmation users' requests to a Web server. A appeal is recorded in a record file entry, which surround dissimilar types of information, as well as the IP address of the computer production the request, the user admittance date and time, the document or image appeal and so on [3]. Depending on the attractiveness of the Web site, a Web log cans confirmation thousands or tens of thousands of requirements every day. To find functional patterns (such as association rules or chronological patterns) from this vast sum of information, requests (or log entries) need to be assembly into usage sessions. A session is defined as a group of requirements made by a single user for a lone steering purpose. A user may have a single session or several sessions during a period of time. Only once these infinitesimal sessions have been identified can common usage patterns among sessions be revealed by Web usage mining algorithms. The most regularly used session discovery method is called timeout, in which a user assembly is usually defined as a progression of requests from the same user such that no two following requests are separated by an interval more than a predefined doorstep. This assembly identification method suffers from the problem that it is complicated to set the time threshold. Different users may have unusual navigation behaviors, and their time intervals between assemblies may be significantly varied. Even for the same user, intermission between assemblies may vary [4]. This paper shows the various draw near of web log methods and its usages for the surfers and the session classification method based on numerical language modeling to database trace logs.

II. LITERATURE REVIEW

Priyanka Patel and Mitixa Parmar [5] Analyze the web servers, log repositories plays a key role as it keeps trace of user pattern for unlike users and thus it is great source of acquaintance. Web Usage Pattern is process of getting the web user browsing example by scrutinize their navigational behavior. An up to date process of classify a web user session is the habitual time spent by user on pages. The time exhausted by user on web site is taken with lofty significance in this paper. To search out the clusters of assembly we applied rough set clustering method. With the amalgamation of statistical result and consequence degree of page the preliminary time out is considered. After that for user session identification session timeout is with dynamism adjusted. Fixed value of congregation time may not give truthful user sessions and so here typical user time spent by user on web site is painstaking. Web user logs contain all details about user activities and truthful user session will give further precise user routing behavior or pattern.

Xiangji Huang · Qingsong Yao · Aijun [6] present an application of a new session detection method based on statistical language representation to database outline logs. Several troubles of the language representation based process are uncovered in the demand, which surround how to decide on standards for the restriction of the language model, how to guesstimate the meticulousness of the session recognition result and how to be trained a language model deficient well-labeled training data. All of these troubles are imperative in the flourishing application of the language representation based method for session identification. We propose answer to these open issues. In particular, new methods for seminal an entropy entrance and the order of the language model are proposed.

José Luis Ortega, Isidro Aguillo [7] their aim is to distinguish the division of number of hits and useless time by web session. It also anticipates finding if there are momentous differences between the length and the length of a session with consider to the point of access–search engine, link or root. Web usage mining was used to determine 17,174 web sessions that were notorious from the webometrics.info web site. Outcome show that both allocation of length and length follow an exponential

decay. Important differences between the different origins of the stopover were also found, being the search engines' users those who useless most time and did more clicks in their assembly We finish that a good SEO policy would be acceptable, because search engines are the principal disinterested party to this web site.

Priyanka Patel et al [8] In the web servers, log repositories plays a key role as it continue record of user prototype for different users and thus it is great source of acquaintance. Web Usage Pattern is process of getting the web user browsing example by analyzing their navigational behavior. A modern process of identifying a web user assemblage is the average time spent by user on pages. The time spent by user on web site is taken with high consequence in this paper. To get the group of session we applied rough set clustering method. With the amalgamation of statistical result and importance degree of page the original time out is calculated. After that for user session classification session timeout is enthusiastically adjusted. Fixed value of session time may not give truthful user sessions and so here usual user time useless by user on web site is measured.

Xiangji Huang et al [9] present a novel session identification method based on statistical language modeling. Unlike ordinary timeout methods, here used a fixed time thresholds for conference identification, we use an information theoretic advance that yields more robust results for identifying session boundaries. The author estimated a new approach by erudition interesting connection rules from the segmented session files. We then evaluate the routine of our advance to three standard conference identification process the usual timeout method, the allusion length method, and the maximal forward allusion method—and find that the statistical language modeling appear generally yields bigger results. However, as with each method, the concerts of this technique are different with changing limitation settings. Therefore, we also evaluate the authority of the two key aspects in our language-modeling–based advance: the choice of smoothing technique and the language model order. We uncover that all standard horizontal techniques, save one, perform well, and that routine is robust to language representation order.

Alam [10] performed the session clustering by applying the Euclidean Distance (ED) measure. The authors conducted the research and put side by side the results with Kmean. While PSO and K-Mean are different in nature and produce different results, the authors did not compare the results with any other kind of PSO based session clustering. In our proposed methodology of preprocessing, we applied the two different types of similarity measures and generated the hierarchical clusters. Khasawneh [11] applied data cleaning method on log files to remove irrelevant entries and to have high accuracy rate. The authors identified the users based on IP Address, date, and time of visit and set of log records visited by the user in that period of time. For session identification, an ontology based algorithm was designed especially for structures of website.

III. ANALYZING WEB LOGS USING VARIOUS TECHNIQUES

A. K-Nearest Neighbor

It is one of the easiest classification techniques. It calculates the distance between various data points on the input vectors and assigns the unlabeled data point to its nearest neighbor class. K is an important parameter. If k=1, then the object is assigned to the class of its nearest neighbor. When value of K is large, then it takes long time for prediction and influence the accuracy by reduces the effect of noise. The KNN is here used to find the attackers with the help of the nearest neighbor, with the help of the nearest parameter it assigned the class and estimate the attackers who will be the intruder and after that it will be intimate to the server [12]. With the help of this method it can analyze the data using the nearest neighbor attributes. It take the data or attribute by the user arisen object, if the user want to search any content in the web means, this technique work by analyzing the nearest person frequently searched object, with the help of the searched object it get the key and if the user of another want to search the same word means it give the related query which already searched by the nearest one. The great advantage is it minimize the session time, disadvantage is nearest neighbor security may get loss because it shows the related search result of nearest one.

B. Decision Trees

Decision tree methods will modernize the manual classification of the documents by produce well-defined queries (true/false) in the form of a tree structure where the nodes correspond to the questions and the leaves make a distinction their corresponding category of the documents. After the tree is perverse, a new document can be effortlessly be classified by situate them in to root node of the tree and run from side to side their uncertainty organization awaiting certain leaf is reached. The advantage of decision trees is that the production tree is easy to tell between even for persons who are not exclusive with the specifics of the model [13, 14].

The club of tree is made by the representation which makes available the user with collective view of the classification logic. A danger of the submission of tree technique is "over fitting" that is if a tree is more than fits then the training data will classifies the training data bad but it would organize the documents to be classify later better. With the help of this technique it solve the web log data by using the two decision, it shows number of results and if we choose any one kind from that means it can go with that and it retrieve the related element or object through this method. It provide better result, less time take, disadvantages is search unwanted items also.

C. Ant Colony Optimization

Ants basically are straightforward being, they jointly forms an ant colony which do significant tasks including shortest path traversal to find food source and information distribution with other ants by producing pheromone. In the field of ant colony optimization, models of collective

astuteness of ants are distorted into useful optimization techniques that discovery uses in computer networking. This algorithm is based on the natural activities of the ant in which an extraordinary substance called pheromone is laid down by the ant who goes out in search of the food and remaining ant follows the pheromone, The shortest path is selected in which greatest pheromone is there as that will be the shortest path as the ant choose the shortest path the ant colony algorithm each ant will preserve a record-set and the traffic that is incoming will be diverted to the path which has the more likelihood as in natural ant which select a path with highest pheromone [15]. The ants work totally in search of new sources of food and concurrently use the existing food sources to shift the food back to the nest. The approach aims at competently avoid the traffic among the nodes and such that the ants never come across a dead end for movements to nodes for building an optimum solution set. Like the ant it go with what the user extract wanted the data. It provides the relevant data to the web user. It provide better result when compare to the above result.

IV. CONCLUSION

Information retrieval is an essential factor for all the people in the world, searching the content in the web if it comes to our relevant content means we can enthuse our self for that information. During our surfing many content were arise, to retrieve the relevant information this paper provides a various techniques of data mining approaches to retrieve easily, in web log, many problems were occurred life timeout because of un wanted information arise, to avoiding this kinds of problem this paper shows data mining techniques to avoid the unwanted data with help of K-NN, Decision tree and Ant colony method, when compared to those method ant colony provide a better search results.

REFERENCES

[1] Tasawar Hussain, Dr. Sohail Asghar, "A Hierarchical Cluster Based Preprocessing Methodology for Web Usage Mining".
[2] José Luis Ortega, Isidro Aguillo, "Differences between web sessions according to the origin of their visits".
[3] Burton, M., & Walther, J. (2001). The value of Web log data in use-based design and testing. Journal of Computer-Mediated Communication, 6(3).
[4] Xiangji Huang, "Dynamic Web Log Session Identification with Statistical Language Models"
[5] Priyanka Patel and Mitixa Parmar "Improve Heuristics for User Session Identification through Web Server Log in Web Usage Mining".
[6] Xiangji Huang · Qingsong Yao · Aijun "An Applying language modeling to session identification from database trace logs"
[7] José Luis Ortega, Isidro Aguillo, "Differences between web sessions according to the origin of their visits".
[8] Priyanka Patel and Mitixa Parmar, "A Review on User Session Identification through Web Server Log".
[9] Xiangji Huang, "Dynamic Web Log Session Identification With Statistical Language Models"
[10] Alam, S., G. Dobbie, et al. (2008). Particle Swarm Optimization Based Clustering Of Web Usage Data. 2008
[11] Khasawneh, N. and C.-C. Chan (2006). Active User-Based and Ontology-Based Web Log Data Preprocessing for Web Usage Mining. Proceedings of the 2006 .
[12] Vikram Singh and Balwinder Saini "An Effective Tokenization Algorithm for Information Retrieval System" CS & IT-CSCP 2014.
[13] Kumar, R. (2009). Mining Web Logs: Applications and Challenges. KDD'09, June 28–July 1, 2009,
[14] Rao, V. V. R. M., D. V. V. Kumari, et al. (2010). "Understanding User Behavior using Web Usage Mining." ©2010
[15] Abraham, A. and V. Ramos (2003). Web Usage Mining Using Artificial Ant Colony Clustering and Linear Genetic Programming. Proc. Of the Congress on Evolutionary Computation (CEC 2003), Canberra, pp. 1384-1391. IEEE.

DOI 10.17148/IJARCCE.2016.5558