Efficient Mining of web log for improving the website using Density Based Spatial Clustering Application with Noise (DBSCAN)

Akiladevi R, Mathew Kurian

Abstract—**Increasing the effective and efficient method for collecting the large amount of data and storing the data and make use of the information contained in the data. To make structure of the dataset called Cluster analysis. The aim is to group the object in a meaningful subclass. It can avoid the manual intervention to select the parameters. Based on the statistical properities of the dataset eps and MinPts values are selected. DBSCAN is less efficient than IDBSCAN. By automatically selecting the parameters, IDBSCAN reduces the execution time. For that, Distance matrix can be calculated by the difference of the session start and end time of each id.**

Index Terms— **Data mining, DBSCAN, Eps, IDBSCAN, MinPts.**

I. INTRODUCTION

To obtain the insight of customer behavior, companies mostly depends upon the websites. There are two types of the logs 1.Server logs and 2. Client logs. Server log can records all the activities on the server. Client log is not used much. The server log contains this following information ip address, session, port, date and time. By using the ip address, each record in the can be uniquely identifiable.

Based on local connectivity and density functions, density based algorithm can be performed. The criteria of each cluster consist of higher number of points than the outside of the cluster. The main feature of DBSCAN is to discover the cluster of arbitrary size and it can handle noise. Epsilion and Minimal points are the two parameters in DBSCAN. The centre point of the cluster is called the core point and all other points except the core point called border point. Consider the point p, if cluster is formed when p is a core point. Continue the process all the cluster is formed.

The input of the DBSCAN is Eps and MinPts can be given manually. Based on the density of the objects, DBSCAN can be performed. By giving the large value of MinPts reduces the number of core objects greatly. By giving the minimum value of epsilion can produces more number of clusters. Core Point is defined as the interior centre point of the cluster. Border Point is the neighbourhood of the core point. Noise point contains the irrelevant data. If MinPts is too large, then small number of cluster is formed. If MinPts is too small, then large number of clusters is formed.

This paper is organized as follows: Section II includes a discussion on key concepts in this paper, Section III gives related works; Section IV and V discuss about the DBSCAN and IDBSCAN. Section VI and VII gives the system model and experimental results. Section VIII gives the conclusion of this paper.

II KEY CONCEPTS

A. Web Log

It is used to analyze the website access log and having the ip address, port, session time. These files are not accessible by normal persons and it can be accessible only by authorized persons. By analyzing the web logs we can tune the sales effort of the specific organization, to improve the websites and redesigning the pages for the website and evaluating the effectiveness of the advertising companies. Common log format and extended log format are the two formats followed by the web logs.

B. Clustering

It is the grouping of similar data. The quality of the cluster depends on the good clustering methods. The single link is the smallest distance of the element within the cluster. The complete link is the largest of the element between the clusters. Based on the similarity, cluster analysis can be done by grouping the objects. Partitioning method, hierarchical method, Density based methods are the clustering methods.

C. IDbscan:

The values of Eps and MinPts are automatically selected based on the statistical properities of the dataset. Distance distribution matrix need to be calculated.

D. Navigational pattern:

The list of pages accessed by the user while browsing the site called the navigational pattern. Suppose, a person can visiting the pages A, B, C, D. The movement of these pages called the navigational pattern.

III RELATED WORKS

There is a huge number of research work done on the web usage mining area to improve the site. Earlier work is based upon the selection of the two parameters Eps and MinPts [1]. Evaluate the effectiveness of the websites by using the web log files. There are two approaches to evaluate the effectiveness of the websites user studies and analyzing the web log data. The time oriented heuristic to identify the unique visitor in the website [2]. There is a method to identify the session. The session is defined as the visiting time of the web pages [3]. The session identification is used
to make the each user requested page into the individual sessions. There are the three steps to analyze the web log data preparation, pattern discovery and pattern analysis [4]. Introduce an algorithm that is suitable for discovering the large databases [5]. To Analyzing the web log for restructuring the websites, improving the sites, monitoring the websites [6]. It is use to cluster the visitors who visiting the similar pages. Data preprocessing is very important from all other process.

IV EFFECTIVE WEB LOG MINING USING DBSCAN

Collect the dataset for the web log analysis. Using the split command split the dataset into appropriate data that contain the information such as ip address, session date and time, protocol, port number. Remove the implicit request and the robot request. From the page view, have to identify the session by using ip address and session date and time. Apply the DBSCAN algorithm, we have to give the value for the two parameter Eps and MinPts and then apply the OPTICS to ordering the dataset and view the cluster. But, we have to choose the parameter carefully, If the Eps value is large , then there will be small number of clusters. If the Eps value is small, there will be large number of clusters. Likewise, MinPts have to be chosen.

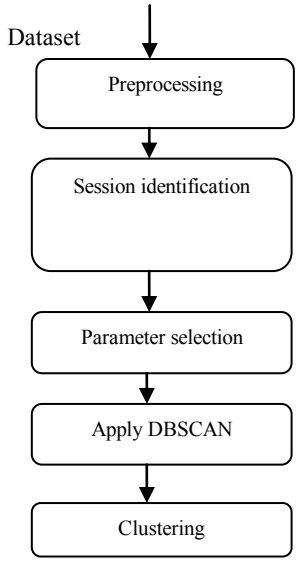

Fig 1. Clustering using DBSCAN

Then Cluster can be formed according to the session id. Inverse document frequency and then predict the online navigational patterns for that kNN approach is used. Term frequency is used to detect the frequent urls in that particular session id and inverse document is used to detect the frequent urls in all session id. The flow diagram for the clustering the urls by using DBSCAN is shown in Fig 1.

V EFFECTIVE WEBLOG MINING USING IDBSCAN

After browsing the web log, the preprocessing, session identification can be performed and instead of DBSCAN, use IDBSCAN. It can avoid the manual intervention to select the parameters Eps and MinPts. For that Distance matrix can be calculated. It can be done by comparing the two sessions dis(i, j) and it can automatically select the Eps and MinPts according to the dataset.

VI SYSTEM MODEL

In this system model, the data in the web log files can be loaded and it undergone for preprocessing and then identify the session and calculate the distance matrix between the session and it can automatically identifies the Eps and MinPts and then apply the IDBSCAN. First, the data undergone for preprocessing and then it can eliminate the implicit request; the result can contain the request that cannot contain the source information. It can identify the session. Session means at what time the user can browse the site and at what time the browser can leave the site. The distance matrix can be calculated by the distance between the sessions and the Eps value can be automatically selected by the mean value of the url and MinPts can be calculated by using the Eps value. Apply IDBSCAN, then it clustering the url.

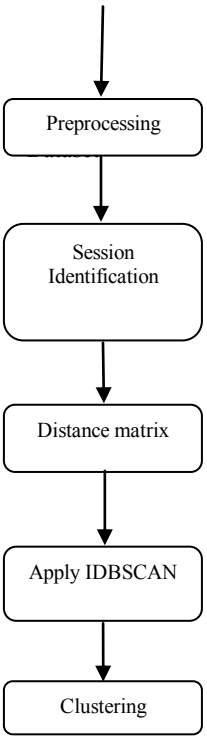

Fig 2. Clustering with IDBSCAN

It mainly used for the information retrieval. TF-IDF is used to determine the relative frequency of words in a specific document compared to the inverse proportion of that word over the entire document corpus. This calculation determines how relevant a given word is in a particular document. TF-IDF value can be calculated by using the formula

$$TF(d,t) = \log\left(1 + \frac{n(d,t)}{n(d)}\right) \quad (1)$$

$$TF.IDF(d,t) = \frac{TF(d,t)}{n(t)} \quad (2)$$

Where n(d, t) is the number of occurrences of term t in document d, n(d) is the number of terms in the document d, n(t) is the number of documents containing term t. In Other words, where n(d, t) is the number of occurrences of pages or websites in the session; n(d) is the number of pages or websites in the session; n(t) is the number of session containing the websites or pages.

Cosine Similarity between two pages or websites can be calculated Bby this formula

$$\cos(A,B) = \frac{A.B}{\|A\|.\|B\|} \quad (3)$$

where A is a query and B is a document. Normalization usually means force all values to fall within a certain range, usually between 0 and 1, inclusive.

VII EXPERIMENTAL RESULTS

The effectiveness of this algorithm is calculated by using accuracy. The accuracy measurement determines how the clustering algorithm is able to create the clusters. Accuracy in the existed system can be improved by using IDBSCAN. The proposed system can be improved by using the IDBSCAN in which the distance matrix need to be calculated. Eps can be calculated by the value of $DIST_{n*n}$. First the distance between session start and the session end of the particular id can be calculated. Suppose if there is 5 session id as shown in table 1

Table 1.Difference between session start and session end time in seconds

Session id	Session start time	Session end time
1	13:31:58	13:32:00
2	13:32:11	13:32:11
3	00:26:09	00:26:23
4	00:03:14	01:09:05
5	00:17:46	00:19:11

For session id 1, the difference between session start and session end time in seconds is 2s. For id 2, difference is 0s. For id 3, the difference is 14s. For id 4, the difference is 3951. For id 5, the difference is 85s. Then the distance matrix can be calculated by using the above values.

Table 2. Distance matrix

S.No	1	2	3	4	5
1	0	2	12	3949	83
2	2	0	14	3951	85
3	12	14	0	3937	71
4	3949	3951	3937	0	3866
5	83	85	71	3866	0

For each DIST (i, j) can be calculated. Then for the DIST (1, 1) can be calculated by using the difference of the session start and end time of all the values. For each i, j is keep increasing. For example I value is 1, then for value 1 we have to calculate the j value for 1, 2, 3 upto 5. The value of DIST(1,1) is 2 it can be subtracted from 2, then DIST (1, 2) is 0 then it can be subtracted from 2 then the value is 0 likewise its processing. The number of objects in the Eps neighbourhood can be calculated. MinPts can be calculated by the mathematical expectation of these objects.

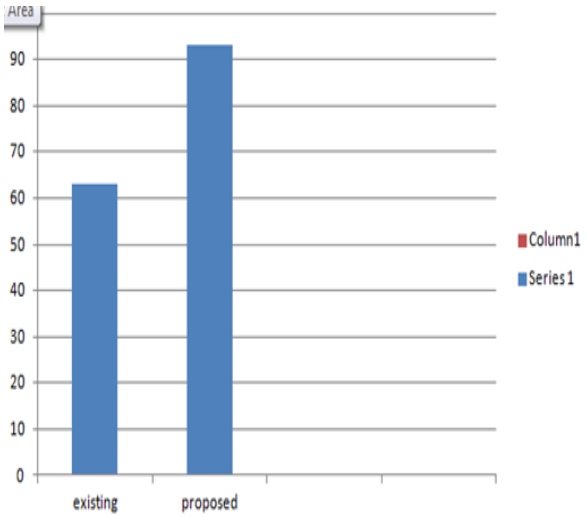

Fig 3. Comparision between existing and proposed system.

VIII CONCLUSION

Use algorithms like IDBSCAN to predict the above activities and it can automatically select the parameters in order to avoid the manual intervention. Distance Matrix is also calculated to select the parameters. There are various benefit of web usage mining, client are targeted with the appropriate advertisement. Also, relevant products in the real time are suggested during browsing the particular website. Web log mining ends by analyzing the results. The enhancement work can be done by using the automatic selection of parameters. So, it will reduce the two step process by one step and increases the accuracy. In future, IDBCAN can compared to other algorithm and find if there is any improvement of accuracy.

REFERENCES

Agosti M, Nunzio G.M.D (2007), "Web log mining: a study of user sessions", in: Proceedings of the 10th DELOS Thematic Workshop.

Berendt B, Mobasher B, Nakagawa M, Spiliopoulou M (2003), "The impact of site structure and user environment on session reconstruction in web usage analysis", Mining Web Data for Discovering Usage Patterns and Profiles 159–179.

Chen M, Park J, Yu P (1996), "Data mining for path traversal patterns in a web environment", in: Proceedings of the 16th International Conference on Distributed Computing Systems, pp. 385–392.

Eirinaki E, Vazirgiannis M (2003), "Web mining for web personalization" , ACM Trans Inter Tech. (3) 1–27.

Ester M, Kriegel H.P, Sander J, Xu X (1996), "A density-based algorithm for discovering clusters in large spatial database with noise" , in: Proceedings of the 2nd Int.Conference on Knowledge Discovery in Databases and DataMining, Portland, Oregon.

Akiladevi is pursuing her M. Tech in Computer Science and Engineering from Karunya University, Tamilnadu, India. She received her Bachelor's degree from Anna University in Computer Science and Engineering.

Mathew Kurian has finished his M.E in computer science and engineering from Jadavpur University, Kolkatta and currently he is working as Assistant Professor in Department of Computer Science and Engineering in Karunya University. Previously, he worked as Software Engineer with Aricent Technologies. He is currently doing his PhD Degree in Data Mining.

