INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

A SURVEY ON WEB LOG MINING USING DBSCAN

Akiladevi R[1], Naveen Sundar[2]

[1]Post Graduate Student, ,Karunya University, Coimbatore
[2]Assistant Professor, Karunya University, Coimbatore
akiladevi001@gmail.com, Naveensundar@karunya.edu

Abstract

Repeated visits and customer retention can lead to the improvement of website. Extracting the data from the web, database called the data mining. Extracting the data from the web log is referred as web log mining. There are several algorithm are used to mine the web log. But only few algorithms can efficiently mine the web log. This study reports on how to predict the navigational pattern and how to mine the web log effectively.

.

Keywords: Data mining, Web log mining, DBSCAN, navigational pattern, kNN ve keywords/phrases are to be provided for indexing purposes.

1. Introduction

Website is the collection of the web pages. it can be accessed by the HTTP (Hypertext Transfer Protocol). It can be accessed from the URL (Uniform Resource Locators). Hyperlink is used to navigate the web pages. Today, data mining techniques are used by many companies to focus the customer retention. Financial, communication and marketing organization are the companies using the data mining techniques. Statistical, machine learning and the neural network are used to analyze the data. Classes, Clusters, association rules, patterns play a vital role in data mining.

Web content mining is used to extract the information like image, text and video, audio from the web. The process of structuring the information from the web is called web structure mining. Web Usage mining is useful for extracting information from the site. This can enable the company productivity through analysis. Web Structure mining is used to analyze and structure the information containing in the website and used for clustering the web pages and then it can scan the data and provide the result based on a requested query. Considering the huge amount of information in the website, there are two problems. Unrelated search results and the next problems is the incapability to index the information in the web. This problem can be reduced by web structure mining. To overcome the above problem, structure mining is used to cluster the information in websites.

Web content Mining is used to mine the information from the web and then integrate the information and review the information to detect the noise. User profiles and the network management are the application of web content mining.

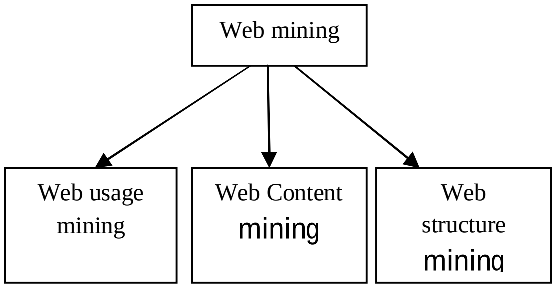

Fig 1. Types of Web Mining

The web log mining is the process of extracting useful information from the web log. This concept belongs to the web usage mining category. It defined as the activities of users when they are in browsing and navigating through the web. Web log contains the information about the frequently visited pages, visits, hits, ip address. The task is to extract the above information from the web. Session is defined as the interval at which the user enters the website and leaving from the site.

1. **Customizing and Optimizing the Websites**

Customization, Optimization and page gathering algorithm is used to improve the website. Customization is the process of adapting the website according to the need of the individual visitors. Facebook is the example for customization; we can customize the facebook profile, facebook background, facebook color, facebook page, facebook cover, and facebook timeline and facebook theme. Optimization is the process to improve the website based upon the interaction with all the visitors. Family vision centre is one of the examples for optimized websites that is used by the eye care centre.

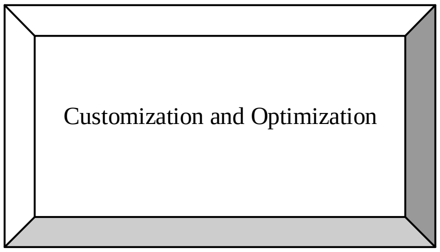

Fig 2. Improving the website

Page gather algorithm is used to find the collection of associated pages in the websites based on the hyperlink map. Parsing, Indexing the document, sorting are the three ways to index the web. A string of symbols can be analyzed by using the parsing analysis. The process of arranging the items called sorting. The way of retrieving the element called the indexing. Input for the page gather is the web log and calculate the frequencies between the pages and the similarity matrix and the graph is created and form the cluster and create the link to the pages. Takes lot of time to calculate the similarity matrix.

The kNN performance can improved by finding the related term assigned to the category [2]. There are two steps to improve the kNN. First step is to calculate the similarity matrix and the next step is to classify the web pages. But the linkage information is not clearly to define the classified pages.

2 Filtering Methods

Visualization can be represented by using the graph. By this, we can restricting the websites, and we can improve the navigation and monitoring the websites. The steps of the web log mining is to collect the data and cleaned the data, user and session are identified, and then select the features and the data can be transformed and combined and then mine the data and at last visualize the result. Cleaning step can varies depending upon the sites more concentration is required otherwise the result is inaccurate.

The process to map the activities of each user into different sessions is done with the help of two activities [4]. They are Proactive mechanism and reactive mechanism. Frame based site has the serious impact. Proactive mechanism is used to provide the correct mapping of each visitor during their activities.

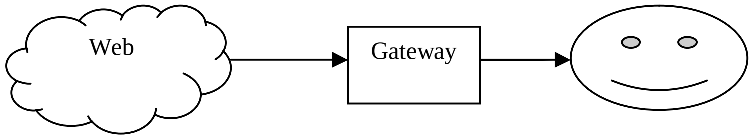

Fig 3. Filtering the users

Identification of cookies and the session are the examples of the proactive mechanism. Reactive mechanism is used to map the activities as posteriori. Page stay time, Page visit time is the examples of the reactive mechanism. It can identify the user with different ip address while browsing the websites. Cookies are not only the possible way for reconstructing the websites. Frame free sites can apply only for small session only.

Access log lists all the requests that requested from the website. Referrer log defines where the visitors come from. A site file contains the files for the site pages. Access log, referrer log and the site files, agent log are the input to identify the user [5].

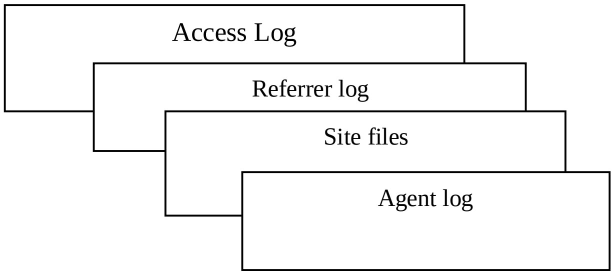

Fig 4. Inputs for the web log mining

Apply the appropriate classification algorithm to classify the pages to filter the sites. Clustering and the association mining algorithms are applied to the log files to analyze the knowledge acquired from the log. User navigational behavior can be analyzed by the maximal forward references.

It is necessary to know the activities that come from the same proxy [6]. Identification of the user does require the knowledge. Distinguishing the users from the same host, proxy is not sufficient. By using the cookies, the user from the same host or proxy can be identified for that cookies identifier is required. To analyze the navigational behavior, entry point and exit point is required. While browsing the website, backward moves are recorded. User can use the link to visit the other pages. In this case, it will produce some problem.

Collaborative based filtering, content based filtering and rule based filtering are approaches for the automatic personalization [7]. Collaborative based filtering can work based upon the human judgments ie., ratings. Based upon the rating, it can filter the content. It I divided into non probabilistic filtering and probabilistic filtering. In non probabilistic filtering, use the user based nearest neighbor item based nearest neighbor and it can reduce the dimensionality. Bayesian network models and the EM algorithms are called the probabilistic models. In these, there are to ratings implicit and explicit ratings. implicit rating is defined as observing the user behavior. Explicit rating is defined as user rating for the pages.

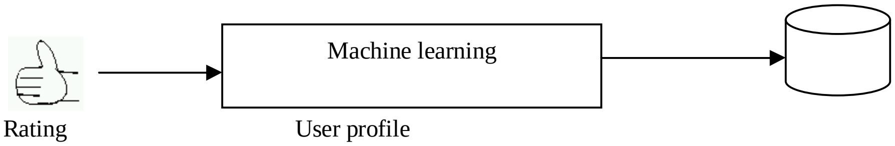

Fig 5. Content based filtering

In content based filtering, recommendations are done by based upon the information in the website. The main problem arises when the content is encoded by the algorithm. The best solution is integrating the content and collaborative based filtering. To predict the navigational pattern in the online can be done by classifying the user navigational patterns [8]. Least common sequence is used to measure the maximum length for both online navigational patterns and the user navigational patterns. Accuracy of the classification can be improved in the online phase.

3. Ordering the Clustering Structure

Ordering the clustering structure is necessary for large data set. Respect to the density based algorithm, we can ordering the clusters [9]. OPTICS can generalize the cluster by ordering the points. So, OPTICS is required to ordering the clusters. OPTICS cannot the point at centre. Reach ability distance is the distance in which all the points become closest to each other point. Single link effect can be avoided by using the density based algorithm. To maintain the algorithm, no index structure is needed..

Outlier detection tool can be used to compare the erroneous data and outliers [10]. It can identify the erroneous data. It needs some tools to detect the outliers. Kmeans algorithm is inefficient to detect the noise. So, the DBSCAN is used to detect the noise efficiently. In Kmeans we have to know the cluster as priori. But in DBSCAN, we need not to know the number of clusters as apriori.

4. Improving the Sites

To improve the websites users requirements are predicted [11]. The preprocessing is the first step. The quality of data depends on the preprocessing. The services and the personal account information are provided by the website. Real time recommendation for personalization is possible.

5. Comparision Table

S.NO	METHODOLOGY	MERITS	DEMERITS
1	Customization and optimization	Protecting site original design from destructive changes	Takes lot of time for gathering the pages to calculate the similarity matrix.
2	Session, Session reconstruction , cookies	To identify users, track their requests, create profiles to statisfy specific needs.	Find pattern among different user with same ip address is difficult
3	Knn approach	Similarity measure is successful to improve the knn.	Not uses linkage information of web pages to classify the web sites .
4	Proactive and reactive mechanism.	It can identify one user accesses the server with different IP combination.	Reactive mechanism is used for reconstructing the session in the presence of cookie is not always possible. It can perform in frame-free sites with small session only.
5	Content based, Collaborative, and rule based filtering	It can manage all the information about the user of the website	Network traffic is there.

6. Conclusion

The survey on DBSCAN can briefly define about the idea of how to improve the website by customization and optimization and then identify the unique users by cookies, filtering approaches. Tracking the user behavior in the proxy level, server level, client level and the maximal forward references is used, ordering the dataset for analyzing the cluster by using the OPTICS. Personalization, monitoring the website by counter measure evaluator.

REFERENCES

Perkowitz M, Etzioni O (1998), “Adaptive websites: automatically synthesizing web pages” , in: Fifteenth National Conference on Artificial Intelligence (AAAI-98)and Tenth Conference on Innovative Applications of Artifical Intelligence (IAAI-98) – Proceedings, pp. 727–732..

Oh-Woog Kwon and Jong-Hyeok Lee, “Web Page Classification Based on k Nearest Neighbor Approach” , ACM 1-58113-300-6/00/009.

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.

Cooley R, Mobasher B, Srivastava J (1999), " Data preparation for mining world wide web browsing patterns" , Knowledge and Information Systems.

Berendt B, Mobasher B, Spiliopoulou M, Wiltshire N (2001), "Measuring the accuracy of sessionizer for web usage analysis", in: Paper presented at the Workshop on Web Mining, First SIAM Internat. Conf. on Data Mining, Chicago.

Jalali M, Mustapha N, Mamat A and Sulaiman N (2000), "**LCS** Based Classification Algorithm for Online Prediction in WUM Recommendation System".

Ankerst M, Breunig M, Kriegel H, Sander J (1999), "OPTICS: ordering points to identify the clustering structure", Sigmod Record 2849–60.

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