Review of Improvement of Web Search Based on Web Log File

Soniya P.Chaudhari [1]
Research Scholar, M. Tech (CSE), PIT, Bhopal, India

Prof. Hitesh Gupta
Professor in Computer Science and Engineering, PCST, Bhopal, India

S. J. Patil
Assistant Professor Department of IT, Bambhori, Jalgaon, India

ABSTRACT

In this paper we review various research of journal paper as Web Searching efficiency improvement. Some important method based on sequential pattern Mining. Some are based on supervised learning or unsupervised learning. And also used for other method such as Fuzzy logic and neural network [5].

Keywords

Clustering, web log, frequent pattern.

1. INTRODUCTION

Web mining is the application of Data mining. Web mining is automatic discovery and extraction of potentially useful and previously unknown information from wed data.web usage mining is described as extracting the information from web. It finds outs what users are looking for on the internet.

Web content mining (WCM) is a bunch of documents connected with links extraction of useful data, information and knowledge from Web page or document. Web document: Web document is any format accessible through the web such as text, images, audio, video, even text has numerous formats.

Web structure mining (WSM): Web structure mining is organization of pages/sites. In Web structure mining graph theory is used to analyze the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided into two kinds:

a) Extracting patterns from hyperlinks in the web. Hyperlink is nothing but structural component that connects web page to different locations.

b) Mining the document structure: It is analysis of the tree like structure of page structure to describe HTML or XML tag usage.

Web usage mining (WUM): In this Data mining is applied to usage information as proxy logs, web server logs. Web usage mining is looking for useful patterns in logs and documents containing history of user's activity. Web usage mining is also known as Web log mining. Figure 1 shows classification of web mining.

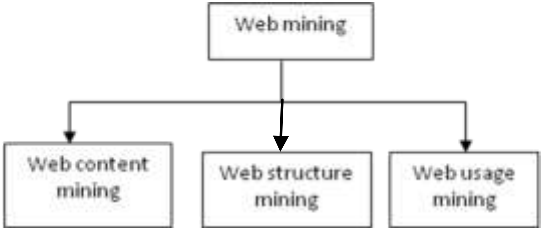

Fig 1: Classification of Web Mining

1.1 Web Search

The web and word web search are essential tools in quest to locate an online information for many people as per there specified query [6]. Web searching is process of finding specified data or information as per users interest. Many aspects are considered during site search. User always tries for finding relevant information as specified by him. So user specifies terms or context associated with his topic. Then using many approach as mentioned in section it finally results related relevant links. Following methodologies are useful for improving web search.

1.2 Cache

In the Internet, proxy servers play the key roles between users and Web sites, which could reduce the response time of user requests and save network bandwidth by buffering frequently accessed documents in the buffer and results into better response time. Cache term indicates temporarily storage of documents that are recently retrieved documents temporary such as HTML pages and images. Web cache is used in various systems as a search engine may cache a website. Web cache is simply storage of web objects allowing user for fast access and helps to improve user experience of the web surfer. Web objects can be cached locally on the user's computer or on a server on the web. There are several caches on user's computer and server on the web [6]. Following are the advantages of web cache:

i) It reduces the load on the web site servers.

ii) It reduces bandwidth needs and cost also which is beneficial for the users, web site owner and service provider.

iii) Faster delivery of web objects to the users.

There are several caches on user's computer and server on the web as follows:

Browser cache: In this caching frequently used web objects in its cache before requesting them from the web site. It uses web objects to increase speed of surfing. Getting objects from cache is faster than getting objects from web site.

Proxy cache: It is installed near the web users. It interprets request coming from a web site and handled by the proxy cache. If they are not present in the cache then proxy gets them from another cache or from website itself.

Transparent proxy cache: Transparent proxy cache intercepts browser web requests without the browser being aware of the interception.

Reverse proxy cache: It is placed in front of the website's server. It reduce the load on a website proxy cache is used called Reverse proxy.

1.3 Frequent Pattern Mining

Frequent pattern mining is proposed by Agrawal in 1993 et al [1]. Frequent pattern mining from the web log optimize the structure of web site as well as it helps to improve the performance of web servers. Aim of discovering frequent in web log data is to obtain information about the navigational behavior of the users. In web log mining different patterns are page sets, page sequences and page graphs.

Frequent pattern mining is beneficial for web users. To mine frequent patterns efficiently many efforts has been done. There are two main frequent pattern mining approaches. Candidate –generation and test approach such as Apriori and its variants, second is pattern-growth and its variants approach such as FP-growth approach.

2. RELATED WORK

Here we discuss different techniques for the improvement of web search and also discuss some other methods related to web log mining for extraction of hit pattern[2]. Many approaches have been suggested to mine information from web access log records from server [3][4].

2.1 WEB Log Mining

Web log is automatically created at server side and it is nothing but a file which saves access pattern of the web user. It is identified extracting web log data. Usually each record in the web log file has the following standard format.

Log name: The remote log name or IP address

Remote host: The remote host name of the user

Date: The date and time of the request

Request: It is exact request line as it came from the user.

Status: The HTTP status code returned to the client

Byte: The content length of the document transferred.

2.2 PREPROCESSING:

The purpose of preprocessing is to extract useful information from web log and transfer it in to desired format. Preprocessing involves three stages as: data cleaning, user identification and session detection. The preprocessing process is shown in fig. 2

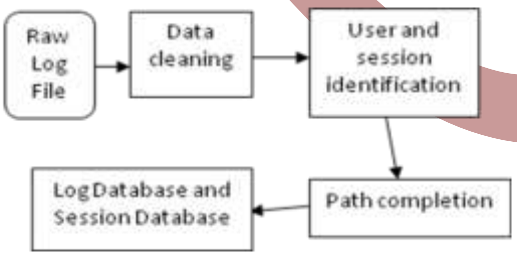
Raw Log File
Data cleaning
User and session identification
Path completion
Log Database and Session Database

Fig. 2: Data Preprocessing Phases

i) Data Cleaning

Web log has lot of unnecessary information which is not directly useful for web log mining procedure. Hence data preprocessing required which cleans redundant log entries and irrelevant data. The entries should be removed by checking the suffix part of URL request. This process cleaned irrelevant information [8] which includes

1. Non-analyzed resources such as images, multimedia files, page style file (Log entry with suffix *. gif, *.jpg, *.css, *.cgi) because these images will auto downloaded with users requested pages.

2. Entries with filename having suffix such as robot.txt.

3. Entries with unsuccessful HTTP status code such as 494(unknown), 404(not found), 501(server error).

4. Remove all entries methods except “GET” method.

5. Response codes ranging from 200-300 are useful.

ii) User Identification

User identification is identification of each user accessing web sites. For user identification some rules are as follows.

1. If there is new IP address, then there is new user.

2. If IP address is same with different browsing software or operating system is assumed that request is coming from different user. The output of user identification gives unique user database which gives information about total number of individual users such as user IP address, browser and user agent.

iii) Session Identification

Session is a sequence of URL’s that are requested by same user (same IP address) within a reasonable time from each other. Session identification divides the accessed pages into umber of sessions. For session identification timeout mechanism is useful by considering following rules [8].

1. If there is new IP address (user) then there is new user.

2. If the refer page is null in one user session then make new session.

3. If time between requested pages exceeds a certain limit (25 or 30 minutes) then we can assume that there is new session.

The outcome of session identification is user IP address along page access for individual user. It gives detail information about total number of session’s as well as session start time, end time details.

2.3 CLUSTERING

Clustering is defined as grouping of data into different sets as per their similarity treads or patterns. Clustering is unsupervised learning meaning that it never needs training data set. It analyzes given data set and observes similarities out of subset of data. Actually the goal of clustering is to identify all sets having similar examples in the data. Instead of diversifying web site search result list, we group the search result into clusters, as user can easily and navigate into particular group that is relevant to user’s interest. There are many clustering algorithms available as discussed bellow.

i) K-means

In k-Means algorithm centroid is the means of web sessions’ similarities in clusters. The summation of variance of current session and each cluster’s centroid is as the criterion function. K-means generates a prototype in terms of centroid, which is usually the means of a group of object data.

1. The K-Means algorithm selects k points as initial cluster centers.

2. In data set each point is assigned to the closed cluster based upon the Euclidian distance between each point and each cluster center.

3. Each cluster center is recomputed as the average of the points in that cluster.

4. Step 2 and 3 repeat until the clusters coverage.

ii) K-medoids

In 1987 this method was proposed for the work with l_1 norm and other distances-medoids chooses data points as (centers) medoids and works with an arbitrary matrix of distances between data points. In K-medoids algorithm we determine the distance between the data point and a representative point of the cluster. K-medoids defines prototype in terms of a medoids and it is the most representative point for a group of objects, and can be applied to wide range of data type since it requires only a proximity measure for a pair of objects.

iii) Clustering with feed forward artificial neural network

As the user sessions are not fixed we recomputed matrix dissimilarities at the time of new session. Therefore this model uses neural network of Olfa Nasraoui and Mrudula Pavuluri introduced in 2004. Then neural network will be used to determine dissimilarities. This model uses feed-forward multilayer perception and it will be trained with back propagation whereas sessions URL's and cluster to prediction model remain in the same [7]. Fig.3 shows simplified view of a feed forward artificial neural network.

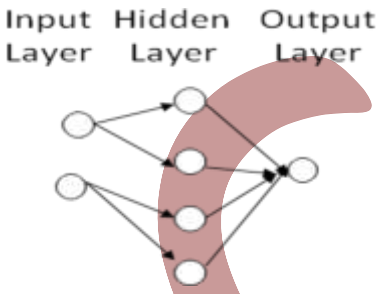

Fig. 3: Simplified Feed Forward Artificial Neural Network

iv) Fuzzy c-Means clustering algorithm:

Fuzzy c-Means algorithm is introduced by Jim Bezdek in 1981.Fuzzy c-means algorithm is much similar as k-means algorithm.

Algorithmic steps for Fuzzy c_means clustering

The Fuzzy c_means attempts set of data points $X=(x_1,x_2\ldots x_n)$ and $C=(c_1,c_2,..c_k)$ be the set of cluster centers.

1. Choose a number of clusters
2. Initialize $U=|u_{ij}|$ matrix U(0)
3. Determine the fuzzy membership' μ_{ij} as

   $\mu_{ij=}$ 1/($(\sum_{k=1}^{c}(dij/dik)_{(2/m-1)}$

4. Calculate the fuzzy centers c_j
5. Stop the iterations when $\max_{ij}(|u_{ij}^{(k+1)} - u_{ij}^{k}|\}<\beta$

Where β is termination criterion between [0, 1] and K is the iteration step.

3. CONCLUSION

Web log is used to improve local search especially the cluster session and neural network. Grouping of web user access session is important to identify web users with similar accessing behavior. In this paper we focus on grouping the session obtained from web log file. Accurate session grouping depends on similarity measures between sessions. In this article we review improvement of web search based on web log file. It presents the data preprocessing phases and web mining techniques. We described some techniques to identify individual users and sessions.

4. REFERENCES

[1] R. Agrawal , T. Imielinski and A. N. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, volume 22(2) of SIGMOD Record, pages 207–216. ACM Press, 1993.

[2] Jin Zhou, Chen Ding, Dimitrios Androutsos (2006), Improving Web site using web server logs.

[3] S. Brin and L. Page, The anatomy of a large scale hyper textual web search engine. In proceedings of the 7th International Conference on computer vision, 1978.

[4] J. Kleinberg, Authoritative source in hyperlinked environment. In proceedings of 9th ACM. SIAM Symposium on Discrete Algorithm, 1998

[5] O. Nasaraoui, H. Frigui, A. Joshi and R. Krishnapuram," mining web access logs using relational Competitive Fuzzy Clustering" to be presented at the 8th international Fuzzy System association world Congress – IFSA 99, August 99.

[6] Fang Yuan, Li-Juan."Study on data preprocessing algorithm in web log mining" Proceedings of the second International Conference On machine Learning and Cybernetics, 2-5 November 2003.