I J C T A, 9(37) 2016, pp. 479-487

Personalized Search Engine Using Web Logs for Educational Institutions

Kushagra Mangal[*], Sagar Kumar Gupta[**] and M. Uma[***]

Abstract: Nowadays, Internet is considered as the best source to attain any information. Search Engines are used to access that information from www. Using a traditional search engine inside the university will not provide always the relevant information for the user. Getting a relevant information from the web is not an easy task, so designing a personalized search system to give more relevant information for a user based on their need is one of the solution to solve this problem. The personalized search result will be based on the ranking of URL by the frequently visited users with in the same network. Hence arises a need for personalized search engine for educational institutions. There are three main techniques of information retrieval (IR), here we propose to use the usage mining techniques and big data analysis to extract the data from the proxy server logs of Institutions network and provide results only in relation to interests of members of the educational institution. The proposed system will analyze the proxy logs, based on the clustering and re-ranking algorithm to give the best result for the search made by the user. This system increases the efficiency and accuracy of search results in figuring out the most relevant information. In the proposed system all the components are going to analyze and give result with the data collected in the same proxy which improves the efficiency of IR than the traditional. This kind of information retrieval will be beneficial for the students and faculty members to get the more relevant, accurate information related to the domain, course material.

Keywords: Personalized Web Search, Usage mining, Clustering, Web Logs, Information Retrieval.

1. INTRODUCTION

Internet is a very useful tool when one needs some information on any topic, but it comprises of huge chunks of data and makes it difficult and time-consuming to extract the relevant information according to user preferences [1]. But most of the common search engine orders their result based on site popularity rather than based on user interests. For example, when keyword like 'cricket' is queried then search results are mainly about the sports game cricket, but it's irrelevant to zoologist. This paper proposes a web search framework to enhance the power of conventional web search.

This paper proposes to use the logs from proxy server of the institution for data mining and personalization of search engine. Proxy servers are used to monitor and limit the bandwidth of data used. By using proxy, there is no need to fetch the same information while that information is still kept on the cache. Furthermore, a proxy server actually stores the list of websites and resources, which were accessed by the users. This log presents the student's and faculty's preference if we could gather appropriate information from the access log. We benefit from this condition. We will use the widely used proxy server to provide user preferences that are stored within its access log. Instead of web crawler that usually wanders across the web to gather the information, this proxy log would also act as information gatherer. The proxy log would be always renewed as the users access the Internet through this proxy so that the information stored in the log would remain updated.

2. PROXY BASED INTERNET ACCESS

Proxy servers are the bridge between the students, faculties requesting data from the internet. Every request from user is first evaluated according to the pre-defined rules and filters the requests and provides the resources by connecting to the relevant server and requesting on behalf of the user [2].

[*] Department of Software Engineering, SRM University, Chennai, India. *Email: kushagra12mangal@gmail.com*
[**] Department of Software Engineering, SRM University, Chennai, India. *Email: sgkmr12@gmail.com*
[***] Department of Software Engineering, SRM University, Chennai, India. *Email: uma.m@ktr.srmuniv.ac.in*

This helps the university in the following manners:

- Giving access to certain educational sites to students connected on network.
- Monitoring the internet traffic of students and the faculties.
- Content filtering according to the predefined rules.

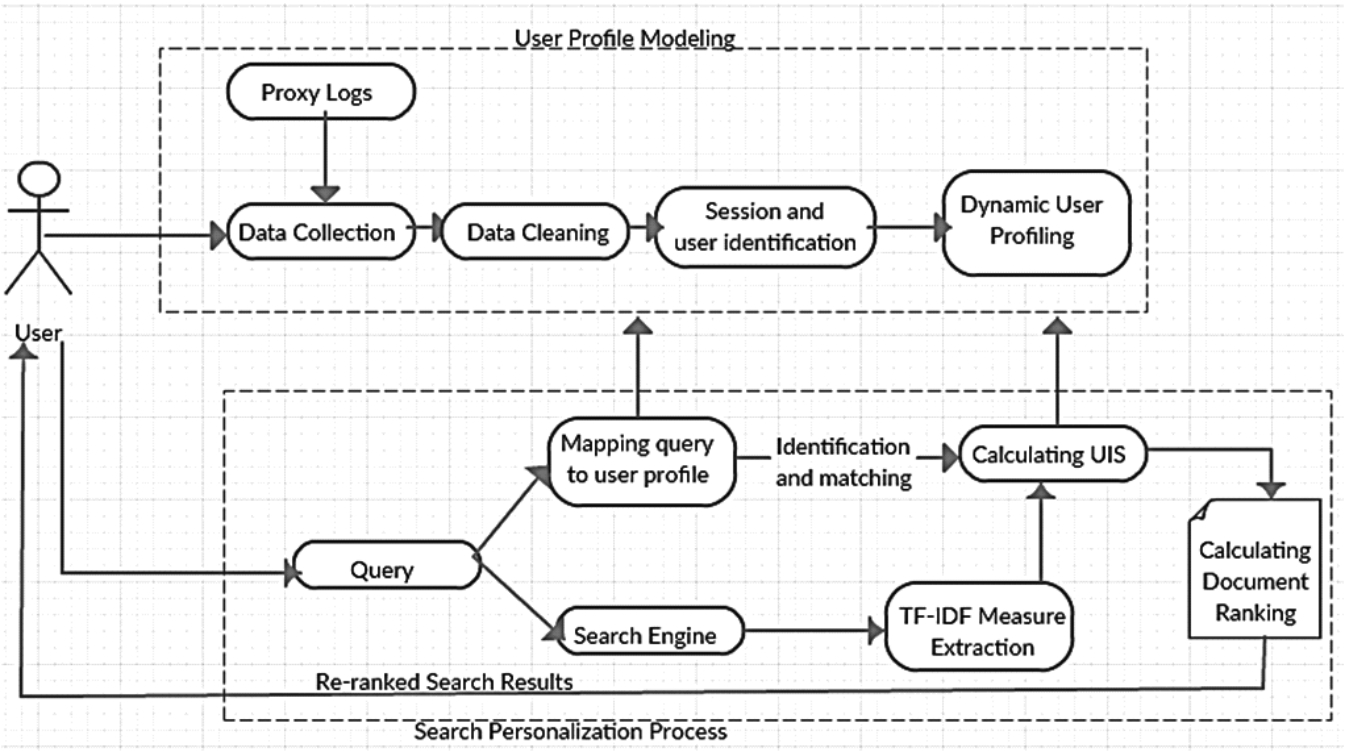

Figure 1: Main Architecture

3. PROPOSED SYSTEM

Our proposed system is based on making a dynamic user profile from the data log on the proxy server and then using that profile to compare with the links extracted from google or bing for the search query user requested, hence re-ranking the pages according to the similarity values that is based on user's interest [3].

Figure 1 shows the architecture of our proposed system which involves the 2 separate steps depending on each other to personalizing the ranking of web pages. These 2 steps are user profile modeling and search personalization process.

In User profile modeling initially the proxy logs are used with the continuous data collection which is then cleaned to remove all the similar links to the pages and the advertisement links from the URLs. Now these filtered URLs are identified and clustered by the k-means method to create a dynamic user profile. This profile shows the popularity of URL according to user's history.

And in search personalization process when user searches a query this query is searched with some other search engine like google or bing, the top k documents are extracted by the TF-IDF method. Also the search query is mapped with the user profile and hence the user index score is calculated on the basis of frequency hits and some other parameters. Now by Ranking this UIS Score with the TF-IDF links the URL are ranked according to the popularity of query in institution, hence personalizing the web search.

4. WEB USAGE MINING

Web mining is a vast field and a rapid growing research area. It consists of Web usage mining, Web structure mining and Web content mining. Web usage mining refers to the discovery of user access patterns from Web usage logs. Web structure mining tries to discover useful knowledge from the structure of hyperlinks. Web content mining aims to extract/mine useful information or knowledge from web page contents.

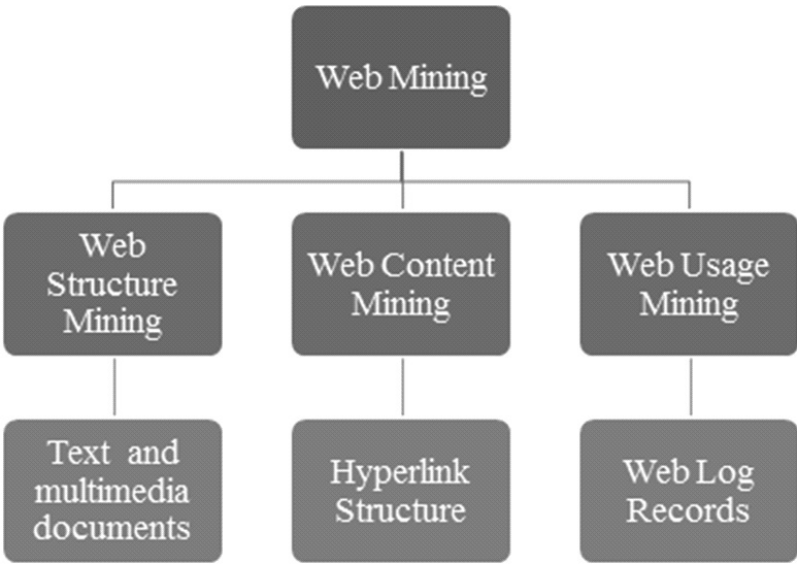

Figure 2: Web Mining Techniques

For personalizing the search engine with proxy logs we will be using Web Usage Mining. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web servers are the common source of data. They store large amounts of information in their log files. These logs generally contain basic information e.g. name and IP of the remote host, date and time of the request etc. The web server stores data regarding request performed by the client. Data can be collected from multiple users on single site. The use of web usage mining in this application helps in retrieval of the data collected from different people over the region that includes-teachers, students and other individual within the campus through the proxy logs and analyzing the data thereafter in order to obtain the access time. Web Usage Mining is also called Web Log Mining or Web Usage Analysis or Click Stream Analysis. Web Usage Mining can further be divide into three different classifications but the one used for the purpose of the collection of the log over the web is Web Server Data. Web usage mining essentially has many advantages which makes this technology attractive even to various government agencies. One of the major flaw or the topic of concern is the issues raised on the privacy, however the use of this technique in an institution as ours does not entertain it. This technique will lead to more of the personalized and accurate study reference for the students.

5. DYNAMIC USER MODELLING

A Dynamic user profile will be crated from the data of clustering and pre-processing. As we are making this search engine as a setup for the university we would be considering university's network as a single user and converting its proxy logs to dynamic user profile. Here dynamic refers to the several users using the university's network and hence it would be updating from time to time. This profile would help identify the popularity of interests in the university. As google considers world as its sample space our sample space will be the members of university's network.

Pre-Processing: It refers to activities done in order to make raw data suitable. It is considered as the first step to mining and analysis of data. Its result is directly used for mining model to obtain the final result. In our proposed system this processed data is clustered and divided into several categories to make the user profile [4].

It involves 4 phases:

- Data Collection: This phase is mainly to collect the data from various sources that may include server logs, client logs etc. and in our case we would be providing with the proxy logs which are collected for the next phase.
- Data Cleaning: Web usage mining basically aims at different user patterns and interests but the log contains several unnecessary entries like advertisement or image links or request entries, which consequently degrades the quality of the use profile hence this process. To clean the data different techniques are used.
- Session Identification: Session refers to series of pages visited by user at particular time. This helps identifying which URL are from which session. The actual reason of this is to create a virtual click stream data for user's actions. This data is then processed finally in next phase.
- Path Completion: This technique is used to recovered missing references of the document. As many times these references are very important from mining point of view. Sometime hits by user are missed by the log due to back button or local cache or proxy server. Hence to solve these missing pieces' path completion method is used.

Clustering: With our proposed system we will be using *k*-means method. It's used for handling and exploring different dataset. Initially it creates *k* groups from the set of objects such that the group members are similar. Clustering analysis is a set of algorithms intended to form groups such that these group members are more similar versus non-group members. In clustering analysis, clusters and groups are synonyms. From the set of vectors, K-means do clustering. The user mentions the number of clusters which are needed. K-means clustering operation has different types of variations for optimizing certain types of data. At a complex level, *k*-means picks the different points and represents them as *k* clusters. These points are called as centroids. Every value will now be closest to one of the centroids. Though they won't be closest to the same one. Hence they will form a cluster around their nearest centroids. Totally '*k*' clusters are formed. Every value is a part of cluster. Now *k*-means finds the center of each cluster based on its cluster member using vectors. This center is new centroid for the cluster. Due to change in the centroid, various values may now be closer to a different centroid. In other words, they may change their cluster membership. Steps 2-6 are repeated such that a point occurs where the centroids no longer shift position and the cluster membership stabilize. This is called convergence. K-means algorithm is supervised or unsupervised. But mostly its classified as unsupervised. We call it unsupervised since *k*-means algorithm learns about clusters by itself without the help of user. User only mentions the number of clusters required. The key point of using *k* means is its simplicity. It's faster and more efficient than other algorithms, especially for large datasets. *k*-means is compatible with Apache Mahout, MATLAB, SAS R, SciPy, Weka, Julia. The advantages of *k*-means are: For huge variables, then K-Means most of the times computationally faster than hierarchical clustering, if we keep *k* smalls. K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular. The limitations are: Difficult to predict K-Value, with global cluster. Another limitation encountered is that different initial partitions can result in different final clusters. It does not work well with clusters (in the original data) of Different size and Different density.

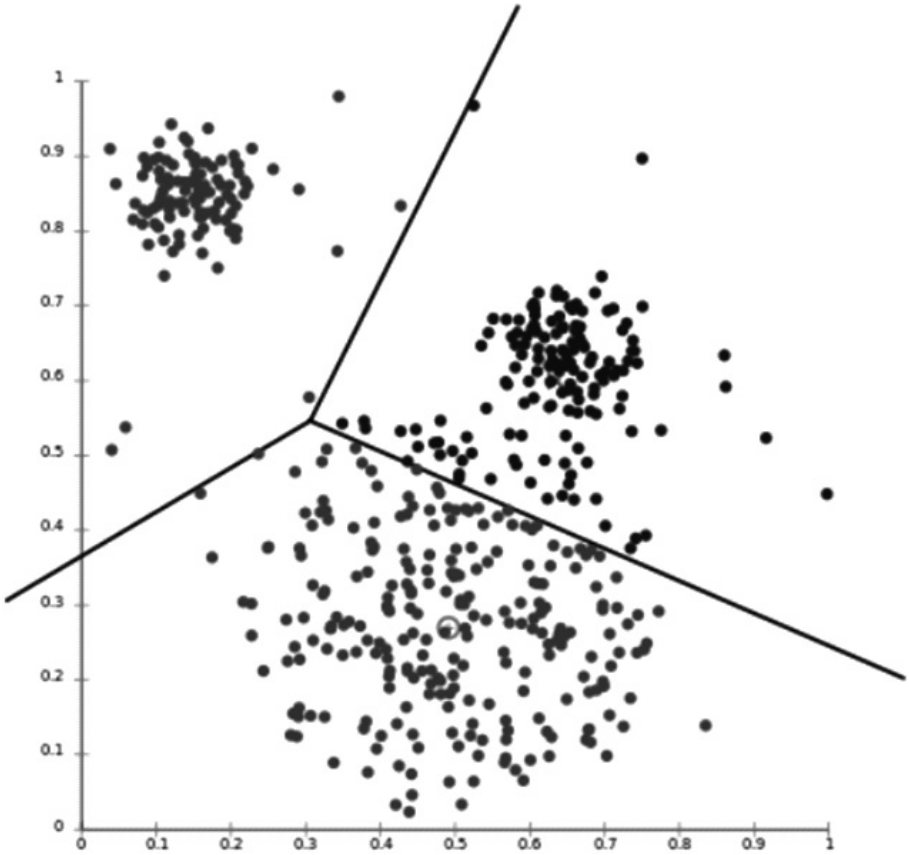

Figure 3: K-means clustering

6. SEARCH PERSONALIZATION PROCESS

The proposed system proceeds through the below processes namely:

- TF-IDF Measure Extraction
- UIS computation

The proposed Personalized Framework process the query for users and provides the personalized or preferred results by weighting the relevant results in accordance with user's interest. When user queries, the search engine retrieves the set of results. From these results, top K documents are selected and these serves as the initial input to the Personalized Network based framework.

TF-IDF measure extraction: TF-IDF computes and measures the top K documents from the server which are analyzed for each term and the same could be retained in the TF-IDF store. From top N terms with high weight are used for further processing that are also stored based on the TF-ICF. The identical terms in every document is collected and their weights are added up and from the outcome, and higher weighted terms are again selected for building the personalized network. Term frequency and Inverse document frequency can be obtained as below Eq. 1: [5]

$$tf_i = \frac{n_i}{\sum kn_k} \tag{1}$$

Where:

n_i = No of occurrence of a term i

n_k = Total no of terms in a document Eq. 2:

$$idf_i = \log \frac{N}{df_i} \tag{2}$$

Where:

N = Total number of documents that are relevant

df_i = Number of documents that contain the term i at least once Eq. 3:

$$\text{TF-IDF weight} = tf_i \times idf_i \tag{3}$$

Thus the term frequency and inverse document frequency are computed.

UIS computation: User Interest Score is computed by considering the different features through which the individual's interest can be tracked. Features are extracted from the Personalized Network and the same are weighted to obtain the UIS.

Features to be considered are:

- Frequency of usage
- Time spent over a concept
- Usage count

User's weight over a concept could be rendered using the top two features and the last feature renders the concept's weight.

The proposed mathematical model computes the UIS and the definitions incorporated are as follows:

User Set U = $\{U_i\}$ where $i = \{1, 2, 3, ..., n\}$

Concept Set C = $\{C_j\}$ where $j = \{1, 2, 3, ..., m\}$:

$C_{ij} = \{F_{ij}, T_{ij}, UC_{ji}\}$

$F_{ij} = \{V\}$

$T_{ij} = \{P\}$

C_{ij} = represents the j^{th} concept for i^{th} user

F_{ij} = represents the frequency of usage of j^{th} concept by i^{th} user

T_{ij} = represents the time spent over the j^{th} concept by i^{th} user

UC_{ji} = represents the usage count of j^{th} concept by all users

Frequency of usage calculates how frequently an individual views a particular concept. Frequently used concept with respect to particular user over a fixed span is computed and it gains the maximum weight among other concepts Eq. 4:

$$F_{ij} = \frac{V_R(C_j)}{\sum V(C)} \tag{4}$$

where, $V_R(C_j)$ corresponds to repeated visits and $\sum V(C)$ corresponds to total number of visits of all concepts over a session.

Time spent over a concept depicts how long a particular concept is viewed by the individual under study. It is obtained by computing the percentage of scroll Eq. 5:

$$T_{ij} = \frac{P_S(C_j)}{\sum P(C_j)} \tag{5}$$

where, $P_S(C_j)$ corresponds to the number of pages scrolled and $\sum P(C_j)$ corresponds to the total number of pages.

Usage count depicts how wide a concept is viewed by various users. This in turn extracts the concept popularity Eq. 6:

$$UC_{ij} = \sum U_i(C_j) \tag{6}$$

where, $\sum U_i(C_j)$ corresponds to the number of users of a concept C_j.

Using the above proposed equaling computation, the higher weighted concept from each user's perspective is obtained. From the higher weighted concept, the weights of the remaining concepts are also calculated relatively. Relative weight is interpreted as below Eq. 7:

$$\text{Wt [Feature(C)]} = \frac{\text{Max [Wt(Fearure)} \times \text{Feature(C)]}}{\text{Max(Features)}} \tag{7}$$

Once features are weighed, the user's interest score of all concepts can be derived using the proposed scoring function Eq. 8:

$$UIS = \sum_{i=1}^{n} \sum_{j=1}^{n} [F_{ij} + T_{ij} + UC_{ij}] \tag{8}$$

Hence the above formula calculates the UIS value.

7. RE-RANKING

For purpose of re-ranking the web search result first the user query is searched on another search engine like google or yahoo then its initial results are extracted and the content and data from those links is stored in a document. This document is later compared with the terms in user's interests and depending on the weightage of the terms the similarity values are defined for the result set and accordingly a new personalized result set is created which would be based on user's interests.

Page ranking: The ranks of the valid results are computed in accordance with the university network's interest. The ranking accounts both TF-IDF measure and user interest score. Personalized page rank is computed as Eq. 9[4]:

$$PPR = 0.55* (UIS) + 0.45* (TF\text{-}IDF) \tag{9}$$

While computing the rank, the weight of the UIS and TF-IDF are varied according to the nature of the query and the user.

8. CONCLUSION AND FUTURE WORK

With the help of this paper we propose a system to utilize the proxy logs in colleges. So far we have not made any comprehensive comparison between our search engine against others. However just proposed a way to generate a web search using web proxy. Figure 4, 5, 6, 7 refers to our sample space which was a single user's history but the same method can be implemented on a larger web log [6].

1311.7388.pdf
Big Data Consulting & Training | Hadoop, Solr and Cascading | Scale U...
Data Mining and Web Mining —
Data Mining Mining World Wide Web
mining - Google Search
Mining - Wikipedia, the free encyclopedia
P2004_4.pdf
Web Data Mining - An Introduction
Web Data Mining, book by Bing Liu
Web Mining
Web mining - Wikipedia, the free encyclopedia
Web Mining and Web Usage Mining Software
Web Mining Tutorial - srivastava-tut-pres.pdf
WEB MINING: A SURVEY OF CURRENT RESEARCH, TECHNIQUES, AND ...
Web Mining: Applications and Techniques: 9781591404149: Computer ...
What is Web Mining? - Definition from Techopedia
What is Web mining? - Definition from WhatIs.com

Figure 4: User search history

Preferred Term	*Frequency Hits*
Web Mining	40
Data Mining	5
Web Usage Mining	15
Mining	3

Figure 5: Terms and frequency hit table

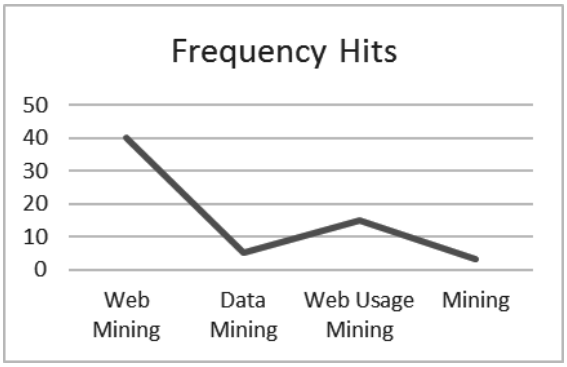

Figure 6: Frequency hit graph

In order to improve this search engine capabilities following future works needs to be carried out:

- Deeper user integration into the search to identify every single user in the proxy.
- Recommendation system to help guide the students and faculties.
- Benchmarking this search engine against an established engine to find its performance.
- Using a wider proxy logs with additional parameters to improve the search results.
- Using a better algorithm to improve its efficiency.

Figure 7: Frequency hit graph

References

1. A. Kumar, and M. Ashraf, "*Efficient technique for personalized web search using users browsing history*", International Conference on Computing, Communication and Automation (ICCCA), Noida, 2015, pp. 919-923.
2. K. Bommepally, T.K. Glisa, J.J. Prakash, S.R. Singh, and H.A. Murthy, "*Internet Activity Analysis Through Proxy Log*", National Conference on Communication, Chennai, 2010, pp.1-5.
3. A. Hawalah, and M. Fasli, "*A Hybrid Re-ranking Algorithm Based on Ontological User Profiles*", 3rd Computer Science and Electronic Engineering Conference (CEEC), Colchester, 2011, pp. 50-55.
4. Dr. S. K. Dwivedi, and B. Rawat, "*A Review Paper on Data Preprocessing: A Critical Phase in Web Usage Mining Process*", International Conference on Green Computing and Internet of Things (ICGCIoT), Noida, 2015, pp. 506-510.
5. Jayanthi, J., and K.S. Jayakumar, "*A Novel Page Ranking Algorithm for a Personalized Web Search*", Journal of Computer Science 8 (7), 2012, pp. 1029-1035.
6. B.S. Hantono, G.D. Putra, "*Generating Customized Web Search Result Through Community Driven Search Engine*", International Conference on Information Technology and Electrical Engineering (ICITEE), 2013, pp. 127-130.