Data Mining in Web Search Engine Optimization and User Assisted Rank Results

Minky Jindal
Institute of Technology and Management
Gurgaon 122017, Haryana, India

Nisha kharb
Institute of Technology and Management
Gurgaon 122017, Haryana, India

ABSTRACT
In the fast moving world, the use of web is been increasing day by day so that the requirement of users relative to web search are also increasing. The content search over the web is one of the important research area comes under the web content mining. According to a traditional search engine, the search is based on the content based matching. But when some site is optimized under the SEO tools, such kind of search is not effective in all ways. The aim of this research is to design a user assisted, reliable, search based on the keyword based analysis ,to provide the user assisted ranked results so that user can select the priority links ,discard the spam links over the web and efficient search optimization model over the open web. The main objective of the work is to implement the work in user friendly environment and analysis of work under different parameters.

Keywords
web pages, data mining, web mining, extreme programming, software tool.

1. INTRODUCTION

1.1 Data Mining
Web usage mining is a subset of web mining operations which itself is a subset of data mining in general. The aim is to use the data and information extracted in web systems in order to reach knowledge of the system itself. Data mining is different from information extraction although they are closely related. To better understand the concepts brief definitions of keywords can be given as [1].

- Data:- "A class of information objects, made up of units of binary code that are intended to be stored, processed, and transmitted by digital computers".
- Information:-"is a set of facts with processing capability added, such as context, relationships to other facts about the same or related objects, implying an increased usefulness. Information provides meaning to data"
- Knowledge :- "is the summation of information into independent concepts and rules that can explain relationships or predict outcomes"

Information extraction is the process of extraction information from data sources whether they are structured, unstructured or semi-structured into structured and computer understandable data formats. Area where data mining is widely used is bioinformatics where very large data about protein structures, networks and genetic material is analyzed. The sub category of interest in this thesis is the web mining which acts on the data made available in the World Wide Web (WWW) data servers.

1.1.1 Web Mining
Web mining consists of a set operations defined on data residing on WWW data servers defines web mining as"…the discovery and analysis of useful information from the World Wide Web". Web mining as a sub category of data mining is fairly recent compared to their areas since the introduction of internet and its widespread usage itself is also recent. However, the incentive to mine the data available on the internet is quite strong. Both the number of users around the world accessing online data and the volume of the data itself motivate the stakeholders of the web sites to consider analyzing the data and user behavior. Web mining is mainly categorized into two subsets namely web content mining and web usage mining. While the content mining approaches focus on the content of single web pages, web usage mining uses server logs that detail the past accesses to the web site data made available to public.

1.1.2 Web Content Mining
"Web content mining describes the automatic search of information resources available on-line". The focus is on the content of web pages themselves. content mining as agent-based approaches; where intelligent web agents such as crawlers autonomously crawl the web and classify data and database approaches; where information retrieval tasks are employed to store web data in databases where data mining process can take place Most web content mining studies have focused on textual and graphical data since the early years of internet mostly featured textual or graphical information. Recent studies started to focus on visual and aural data such as sound and video content too [2,3].

1.1.3 Web Usage Mining
The main topic of this thesis is the web usage mining. Usage mining as the name implies focus on how the users of websites interact with web site, the web pages visited, the order of visit, timestamps of visits and durations of them. The main source of data for the web usage mining is the server logs which log each visit to each web page with possibly IP, referrer, time, browser and accessed page link. Although many areas and applications can be cited where usage mining is useful, it can be said the main idea behind web usage mining is to let users of a web site to use it with ease efficiently, predict and recommend parts of the web site to user based on their and previous users actions on the web site[4].

1.2 LITERATURE SURVEY
In Year 2011, D. Choi has defined an approach to perform the query over the web and to extract the web document. The author also presented the approach to assign the ranking to these web documents. With the development of web search engines, one of the major tasks is to retrieve these documents from web effectively. These search engines uses the some ranking algorithm to present the result in an effective way.

The author has defined a study of existing ranking algorithm used by different search engines. The author explored the advantages and limitations of these ranking algorithms. The major contribution of author was the definition of query based information retrieval. The author defined the classification over a query and performed the query filtration. Based on this analysis, the ranking is improved and refined [5]

Zhou Hui[6] has presented a work on optimization of search engine under the keyword analysis along with face link analysis and back link analysis. Author defined a relational environment based on search engine optimization os that the search ranking will be improved. Author also discussed various aspects of search engine optimization including the optimization vector, ranking, working principal etc.

Ping-Tsai Chun[7] has presented a search engine optimization approach under the current market scenario analysis. Author defined the web service analysis to improve the business dictating and to provide the work under small organizations so that the effective keyword analysis based search will be performed. Author presented the work for text search as well as for image search over the web. The pattern analysis is defined to perform the effective search over web.

One of the common model for web page ranking and prediction system is defined by Markov Model. Such model defines the navigational behavior of web graph theory as well as defines the transitional probabilities over the ranking analysis. The author not only defined work for a single web page access, but also presents the work for web path generation. The web path is actually defined as a series of web pages that a user can visit after visiting a specific web page. To perform this kind of analysis a Markov Model based prediction system is defined. The prediction is here defined under the web usage mining that defined the structural information for prediction of web pages. To perform such kind of analysis, the author defined a web page graph and implements the markov model over it to analyze the frequency match. Based on which a acyclic web path is generated and based on the weightage assigned to this web path the prediction is performed [8]. Another work in web page ranking is the comparison of different web pages and the web sites. Author M. Klein performed this comparison on two football team web sites of college team. The analysis is performed under the web page metrics to perform the quality assessment. The author has defined the page comparison and the ranking system under the graph theory[9].

2. PROPOSED APPROACH

The proposed work is about to optimize the topic based Web Service crawling process with the concept of exclusion of duplicate pages. For this a new architecture is proposed, this architecture will use the rank based service selection approach. In this work the ranking is performed respective main criteria's called User Interest Analysis. The user interest analysis is further categorized under three sub categories called User Query Relevancy Analysis, User Recommendation to service in terms of like dislike factor and the user Web service visits. The basic phenomenon is given as under.

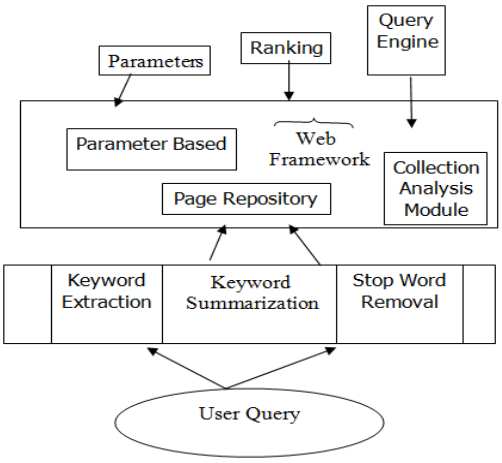

Figure 1: Proposed Web Search Architecture

As we can see in this proposed architecture the user will interact to the Web Service with his topic based query to retrieve the Web Service pages. As the page is query performed it will perform request to the Web Service and generate the basic url list. Now it will retrieve the data from the Web Service. For the url collection it will use some concepts like indexing and the ranking. The indexing will provide a fast access to the Web Service page where as ranking will arrange the list according to the priority. Now as a Web Service page is fetched, the proposed approach will retrieve the keywords form the document and perform the relevancy match by performing the match of service keywords with user query. Now as a new page is retrieved it will generate the suffix tree and perform a suffix tree based comparison to analyze the relevancy ratio. Based on this factor the initial ranking is assigned to the Web service. Here Fig 1 is showing the proposed web architecture. The web architecture is divided in three mains stages. At the earlier stage, the keyword analysis is performed in terms of keyword identification over the query and removes the stop list words from the query. After this stage, the keyword extraction over the query is obtained. Now this keyword list is considered as the query and passed over the web. As the web contents are extracted over the web, the link analysis is performed. This analysis will obtain the web contents and perform the content based match to obtain the most relevant pages over the web.

The steps involved proposed work is presented in the Figure 2.

Crawling the Web Page based on query
Parsing the Web Contents to Text Form
Identify the relevancy Vector and assign initial ranking
Obtain User response
Estimate ranking
Result will be displayed

Figure 2: Process Model of Proposed Work

3. PROPOSED ALGORITHM

Algorithm {

1. Initialize the Web Environment
2. Get user Query
3. Accept the user query and filter it to retrieve the keywords under the following step
    a. Remove the stoplist words from the query list
    b. Remove the Similar Words
    c. Extract the Keywords From the Query
4. Use these extracted keywords as the main query to the web system
5. Extract the web contents and find the occurrence of the keywords in the web pages
6. Find the maximum match web page from the web respective to its contents as well as internal link contents
7. Find the list of M web pages from the web that satisfy the relevancy vector
8. For i=1 to M
   [Perform the Content based similarity measure as]
   {
9. RelevancyVector = 0
   For j=1 to Length (UserKeywords)
   {
   RelevancyVector=RelevancyVector +
   KeywordOccurance(Page(i) ,Keyword(j))

   /TotalKeywords(Page(i),Keyword(j));
   }
10. Check the Existence of Particular web server in Database if it does not exist then set this relevancy vector as the initial ranking parameter
11. Obtain the rank based on user response parameters i.e. like, dislike and
12. Update the rank based on user response.
    }
13. Show the ranked list of pages to user

}

4. EXPERIMENTAL RESULTS

Here Figure 3 is showing the graphical screen on which the user pass the query to search engine. The results are obtained for the web user based on this query by performs server side check under different parameters.

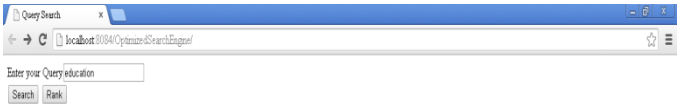

Figure 3: Graphical Screen

Here Figure 4 is showing the results obtained from the web server based on user query. The query results are presented as the base results and decide the initial ranking based on the relevancy vector,

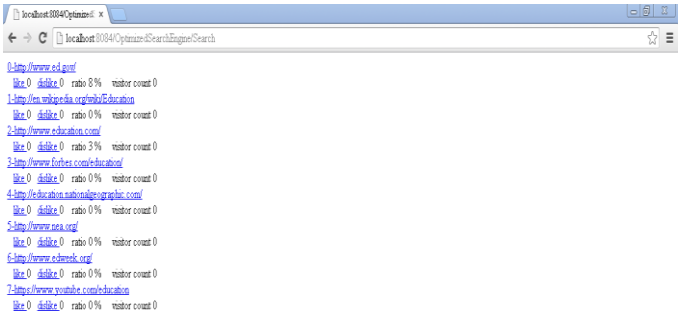

Figure 4: Initial Results

The given Figure 5 is showing the results based on proposed user assisted weighted model. The ranks to links are assigned. The primary ranking is based on relevancy vector. As the like vector will be improved, the rank will be improved and as the dislike vector will be increased, the rank will be decreased

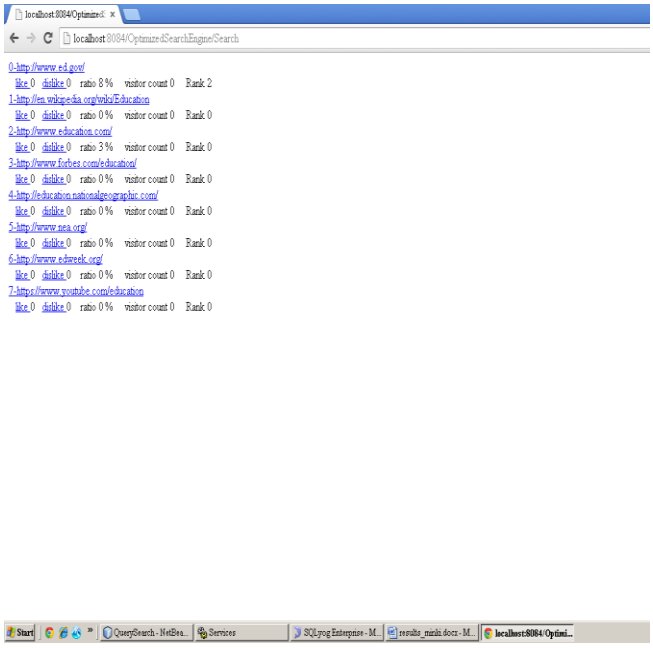

Figure 5: Ranked Results

The given Figure 6 is the impact of like vector. The like clicks on web link "education.com" is increased, ranking of the particular link is increased.

Figure 6: Modified Ranked Results Based On Response

Here figure 7 is showing the analysis of crawled pages under different parameters based on which the ranking to the web pages is assigned. These parameters includes like, dislike, visit count and ranking. Here figure is showing, As the user response is provided to these pages, the ranking is changed. As shown in figure, As the like vector is increased, the ranking of particular page is also increased. In same way, dislike vector and visit count also affect the ranking.

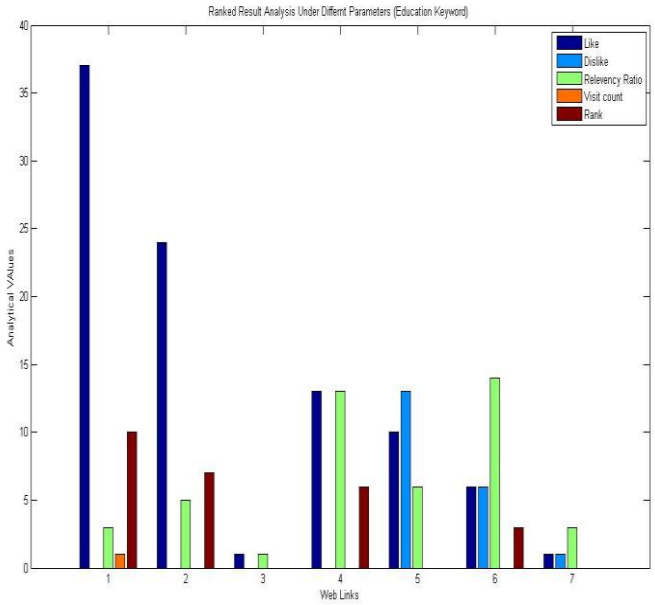

Figure 7: Ranked Page Analysis for Education Keyword

Here figure 8 is showing the analysis of crawled pages under different parameters based on which the ranking to the web pages is assigned. These parameters includes like, dislike, visit count and ranking. Here figure is showing, As the user response is provided to these pages, the ranking is changed. As shown in figure, As the like vector is increased, the ranking of particular page is also increased. In same way, dislike vector and visit count also affect the ranking.

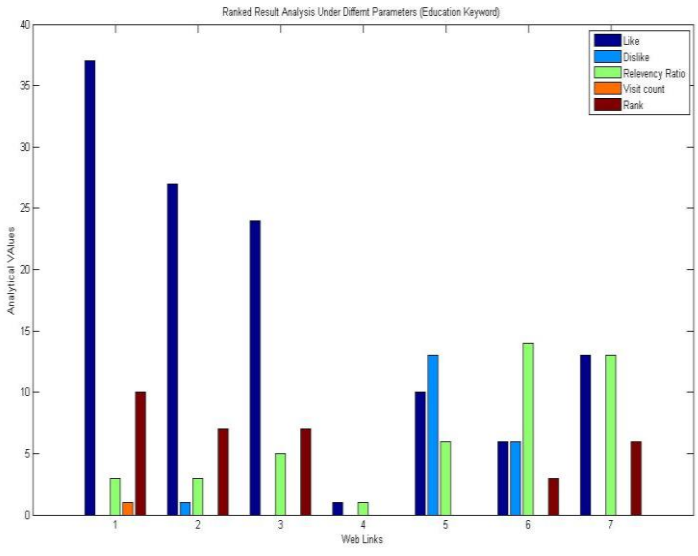

Figure 8: Ranked Page Analysis for Education Keyword

5. CONCLUSION

In this paper we have mainly presented work is about to perform the effective search in the Web environment based on the user query relevancy factor. The relevancy of the query is here analyzed under three main factors called Keyword based Analysis, User Recommendation Analysis and the User Web service visit analysis. Based on these all factors a ranking criterion is decided and based on these ranking vectors the Web services are ordered. The user can get the best Web

service as well as recommend other for the best service selection. In this work, the google App is used as the public Web repository to perform the query analysis. The work is implemented in a web environment to perform the user query and to derive the ordered results from the query.

6. REFERENCES

[1] Rajeev Motwani," Evolution of Page Popularity under Random Web Graph Models", PODS'06, June 26–28, 2006, Chicago, Illinois, USA. ACM 1-59593-318-2/06/0006.

[2] Ravi Kumar," Rank Quantization", WSDM'13, February 4–8, 2013, Rome, Italy,ACM 978-1-4503-1869-3/13/02.

[3] Paul Alexandru Chirita," Using ODP Metadata to Personalize Search", SIGIR'05, August 15–19, 2005, Salvador, Brazil. ACM 1-59593-034-5/05/0008.

[4] Ricardo BaezaYates,"Web PageRanking usingLink Attributes",WWW2004, May 17–22, 2004,NewYork, USA.ACM 1-58113-912-8/04/0005.

[5] Donjung Choi," An Approach to Use Query-related Web Context on Document Ranking", ICUIMC '11, February 21–23, 2011, Seoul, Korea. ACM 978-1-4503-0571-6.

[6] Zhou Hui, Qin Shigang, Liu Jinhua, Chen Jianli, "Study on Website Search Engine Optimization", International Conference on Computer Science and Service System, pp 930-933, 2012.

[7] Ping-Tsai Chung, "A Web Server Design Using Search Engine Optimization Techniques for Web Intelligence for Small Organizations", Proceedings of IEEE Conference, pp 1-6, 2013.

[8] Magdalini Eirinaki," Web Path Recommendations based on Page Ranking and Markov Models", WIDM'05, November 5, 2005, Bremen, Germany ACM 1-59593-194-5/05/0011.

[9] Martin Klein," Comparing the Performance of US College Football Teams in the Web and on the Field", HT'09, June 29–July 1, 2009, Torino, Italy. ACM 978-1-60558-486-7/09/06.

[10] JOHN B. KILLORAN, "How to Use Search Engine Optimization Techniques to Increase Website Visibility", IEEE TRANSACTIONS ON PROFESSIONAL COMMUNICATION, VOL. 56, NO. 1, pp 50-66, MARCH 2013.

[11] Chen Wang," Extracting Search-Focused Key N-Grams for Relevance Ranking in Web Search", WSDM'12, February 8–12, 2012, Seattle, Washington, USA. ACM 978-1-4503-0747-5/12/02.

[12] Bin Gao," Semi-Supervised Ranking on Very Large Graphs with Rich Metadata", KDD'11, August 21–24, 2011, San Diego, California, USA. ACM 978-1-4503-0813-7/11/08.