Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 4, April 2014, pg.707 – 710

RESEARCH ARTICLE

Hybrid Approach for Optimizing the Search Engine Result

Ashish Kumar Kushwaha[1], Nitin Chopde[2]

[1]ME (CSE 1st year) GHRCEMA, India

[2]HOD (CSE) GHRCEMA, India

[1] kushwaha_ashish.ghrcemamecse@raisoni.net; [2] nitin.chopde@raisoni.net

Abstract— ***Due to tremendous growth in growth of internet over recent years, huge amount of data collected over the web and search engine users facing problem in search a relevant information by writing few keywords, search engine returns a number of result page and then user have to spend long time to search a relevant information from number of result. In this paper, we propose a hybrid approach for optimizing the search engine results using document clustering, genetic algorithm and Query Recommendation to provide the user with the most relevant pages to the search query. This process starts with query recommendation, based on learning from query logs that predicts user information requirements in which an algorithm has been applied to recommend related queries to a query submitted by user and process of document clustering, genetic algorithm are applied to resultant pages from query recommendation to deliver most relevant result to user at minimum time.***

Keywords— ***Document Clustering; Genetic Algorithm; search engine; Query Recommendation; Hybrid model***

I. INTRODUCTION

Due to the tremendous growth of the Internet in recent years, huge amount of data is added to the World Wide Web, search engines have to perform complex task of sorting billions of pages and displaying only the most convenient and relevant pages for the submitted search query. With this huge amount of data over on web lead to difficulty in managing and displaying data according to end user perspective and become bottleneck for SEO Engineer and Webmaster. It becomes very essential to promote a website in search engine result in website development. Webmaster or search engine optimization engineer have to be actively learning the techniques and algorithms that drive visitors to their site. For this purpose some ordering of webpage is in result list became important. Most relevant page should be place on the top of list and least relevant page should be at bottom

according to user query. For this purpose ranking of web page is needed for arranging of page according to user demand dynamically. Page ranking is assigning a value (rank) to the web page among the similar type of page to decide its importance. In this we present some algorithm used in page ranking and their comparison and work proposed aims to optimize the results of a search engine by displaying the more relevant and most user relevant pages on the top of search result list. For this we propose a Hybrid of Query Recommendation and Document clustering, Genetic algorithm. This approach starts with finding most popular query by pre-mining the query logs to fetch the potential clusters of queries and from this all clusters we get most popular queries. Every cluster entries are again mined to obtain sequential patterns of pages accessed by the users. After both mining process, output of both mining is combined to get relevant pages to users with recommendation of popular historical queries. After this document clustering and genetic algorithm is applied resultant output. Document clustering is applied to output to group all similar pages together in one cluster (partition) after genetic algorithm is applied on results to optimize the result and Select the best pages which have highest score depending on other features like number of keywords. At last list of web pages are chosen from different regions of information which are the result of genetic algorithm. This give a optimize list of WebPages for user demand query in a short time.

II. RELATED WORK

There is several research studies have been worked on this topic and a good number of existing research paper are available for query recommendation and document clustering, genetic algorithm. The use of Web logs to develop different aspects of search engines studied by R. Baeza-Yates [1].A technique for clustering the similar queries to recommend URLs to frequently asked queries of a search engine has been suggested by J.Wen Nie and H.Zhang [3].A new way to determine the related queries based on association rules offered by B.M Fonseca, P.B Golger [2]. In this paper the queries represent items in tradition association rules. The query log file is viewed as set transactions that represent a session at which the user submit all related queries in a particular time. Evaluating document clustering for interactive information retrieval [4] by A.Leuski suggest a clustering has been used for helping the user in browsing a collection of documents or in organizing the results returned by a search engine, or by a meta-search engine in response to a user query. As discussed in [4], the use of clustering in information retrieval (IR) is based mostly on the cluster hypothesis: ―closely associated documents tend to be relevant to the same request. In [7] the authors have discussed a new way of combining the clustering and genetic optimization in improving the retrieval of search engine results in different settings it is conceivable to design search methods that operate on a thematic database of web pages that refer to a common body of knowledge or to specific sets of users. They have considered such premises to design and develop a search method that deploys data mining and optimization techniques to provide a more significant and restricted set of pages as the final result of a user search. They adopt a vectorization method based on search context and user profile to apply clustering techniques that are then refined by a specially designed genetic algorithm.

The proposed query recommendations [6] concept works on query log which consist of attributes like query name, rank, time clicked URL. Then, the similarity based on keywords as well as clicked URL's is calculated then clusters have been obtained by combining the similarities of both clicked URL's and keywords to perform query clustering and using the modified version of an existing sequential pattern mining technique the sequential order of clicked URLs in each cluster has been discovered. The resultant output is optimized by re-ranking the pages using the discovered sequential patterns.

The proposed method of Intelligent Model [5] uses the Meta-data that is coming from the user preferences or the search engine query log files. These data is important to find the most related information to the user while searching the web.an approach to reduce the results in a short list by applying artificial intelligence techniques such as document clustering and genetic algorithm [5]. Document clustering is required to group all similar pages together into one partition (cluster) after that optimization of the results is applied using genetic algorithm to select from each cluster the best pages with high scores and other features like number of keywords. Finally

the outcome of the genetic algorithm is the final shortlist of web pages that are chosen from different regions of information. Thus we have a reduced number of web pages that can be reviewed by the users in a short time.

III. PROPOSED WORK

The proposed hybrid model (fig. 1) is the hybrid of Query Recommendation and *document clustering, genetic algorithm, model consists of* Query Recommendation system in paper [6] learning from historical query logs. This proposed system calculate user's information requirements in a better way by performing query clustering to find the similarities between the two queries, which is based on user query keywords and clicked URLs. After that Generalized Sequential Patterns algorithm is used to generate the frequent sequential pattern of web pages visited by user in each cluster then previously assigned rank score of the web page are modified to re-rank the search result list by using the discovered sequential patterns. The relevancy of the web pages based on its access history is enhanced by rank updation.

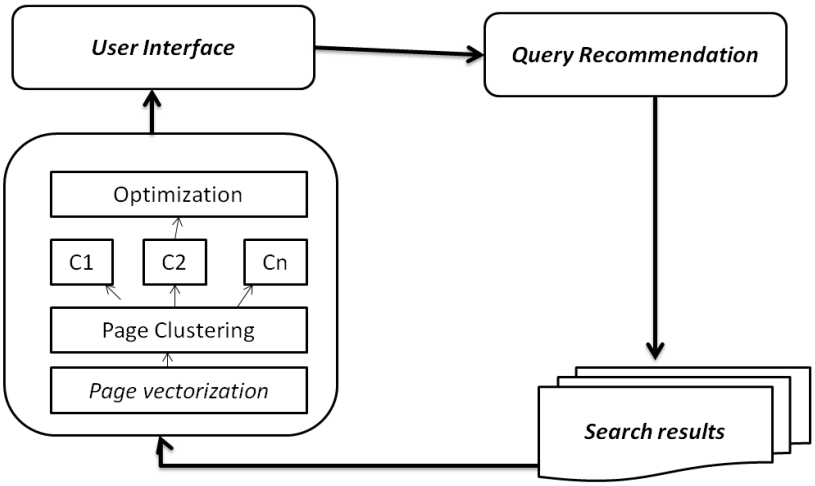

Fig 1. Proposed Hybrid model

After that, the frequent sequential patterns of web pages visited by the users in each cluster are generated with the help of Generalized Sequential Patterns algorithm. The final approach is to re-rank the search result list by modifying the previously assigned rank score of the web pages using the discovered sequential patterns. The rank updation enhances the relevancy of the web pages based on its access history. By this method, the time user spends looking for the required information from search result list can be reduced and the more relevant Web pages can be obtained.

The proposed architecture of Query Recommendation system in paper which consists of following functional components:

- Query Log
- Query Similarity
- Query Clustering Tool
- Sequential Pattern Generator
- Rank Updater

When user writes a query on the interface of search engine, query terms are matched with the index repository of the search engine by query processor and produce a list of matched document. Result optimization system performs its task of gathering user intentions from the query logs in reverse order. Query similarity module continuously analyse the user browsing behaviour as well as the submitted queries and clicked URLs get stored in the logs. The output of which is forwarded to the Query Clustering Tool to create potential groups of queries based on their similarities. Sequential patterns of web pages in every cluster are discovered by Pattern Generator module. Matched documents retrieved by query processor are input to Pattern Generator module. Sequential patterns improve the rank of page which contains search context and the user preference. This improved ranked list is feed to Intelligent Search Engine described in paper [5]. In this first step is Page vectorization in which list from sequential pattern is used to create vector of characteristics for each page. Then these vectorized pages are clustered into similar page called cluster this step is known as page clustering. Finally in third step optimizing is done by applying genetic algorithm on structure identified by the cluster and the score of pages for selecting the best sets of page from each cluster to get most relevant result for user demand query.

IV. CONCLUSION AND FUTURE WORK

In this paper, proposed hybrid approach of optimizing using query recommendation and Document clustering, genetic algorithms can be useful for search engine to optimize the displaying result and able to display the most relevant WebPages with recommendation to user query so user not have to search through list of displayed page and seeking time the of user to retrieve the needed information from the list of pages is reduced by displaying most relevant and use information at the top as per user requirement.

In future, query clustering and page clustering will be combined for as a single phase so the time for both clusters will be minimizes and we will able to provide the most relevant in least time.

REFERENCES

[1]. R. Baeza-Yates, Web Usage Mining in Search Engines." Web mining: applications and techniques, Anthony scime, Editor, Idea Group. 2004.

[2]. B.M Fonseca, P.B Golger, E.S. De. Moura and N.Ziviani. " Using Association rules to discover search engines relating queries". In first Latin American web Congress, November, 2003.

[3]. J.Wen Nie and H.Zhang, Clustering user queries of a Search Engine. In Proceedings at 10th international World Wide Web Conference, pp 162-168, W3C, 2001.

[4]A.Leuski, "Evaluating document clustering for interactive information retrieval.," in Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management, Atlanta, Georgia, USA, 2001, pp. 33–44.

[5] H.M.Zahera, G. F. El Haddy ,A.E. Keshk, "Optimizing Search Engine Result using an Intelligent Model"2012

[6] N.Taneja, R Chaudhary, "Query Recommendation for Optimizing the Search Engine Results" 2012.

[7] M Caramia, G Felici, and A Pezzoli, "Improving search results with data mining in a thematic search engine," Computers & Operations Research, pp. 2387–2404, 2004.

