I J C T A, 10(8), 2017, pp. 91-97 ISSN: 0974-5572

User Dependent Query Recommendation and Search Optimization for Information Retrieval

Aniket M. Akarte* and Pradnya V. Kulkarni**

ABSTRACT

The process of the finding relevant and exact information from the web (World Wide Web) is the most difficult task due to continues growth in the data over internet; along with that some webpages contain malicious data that harms user search, to solve such type of problem in web searching. This paper mainly focus on user personalization based document retrieval, information extraction and queries recommendation to get exact, and accurate information according to the user profile or hobby, we use time dependent analysis and genetic algorithm with cosine similarity fitness function to improve searching efficiency, by the use of the query recommendation module in our proposed system, user can get fresh and more precise data as compare to traditional system of information retrieval , it will also helpful for user future search and saves user operational time, proposed system also able to solve the issues of ambiguous query processing by the help of deduce sensing mechanism.

Index Terms: Web Mining, Genetic Algorithm, Cosine Similarity, User Logs, Information Retrieval.

1. INTRODUCTION

With the continues growth of web documents, finding exact information and extracting information effectively had become most difficult task .User personalization based on data retrieval and query recommendation is the basic aim of this project. There are many ways for user personalization based document retrieval, such as search history based ,deduce sense mechanism based, clustering based on query flow graph and so on explained in next section .These methods of clustering have major drawback because temporal and textual features are considered. Although the time-dependent and in some cases text-based relevance matrix may work well.

For time based matrix, one can assume that a topic related queries has always followed by a related group of queries; however, this may not be the case when the user was multi tasking. Similarly, the text-based metrics such as jaccard similarity and cosine similarity can capture the relevance between query groups around textually similar queries such as apple fruit and apple iPod but such system fail to identify relevant queries and groups of similar queries such as iPod and apple fruit since they are not logically similar. Additionally, the text-based metrics may mistakenly identify query groups around, say, jaguar car manufacturer and jaguar animal reserve as relevant, since they share some common text.

Therefore, we will need a relevance measure that is robust enough to identify similar query groups beyond the approaches that simply rely on the textual content of queries or time interval between them, Proposed system use genetic algorithm with cosine similarity fitness function to identify relevancies between query's and documents then document ranking is to be perform accordingly, also queries recommendation module in this project we will able to determine the relevancy between query's for future search more effectively.

* Dept. of Comp. Engg. MIT Pune, *Email: a1.aniket@hotmail.com*

** Dept. of Comp. Engg. MIT Pune, *Email: pradnya.kulkarni@mitpune.edu.in*

In addition user search history of a large number of users contains matrix regarding query relevance to find exact data according to the user, such as queries tends to be issued closely together, and which queries tends to lead to click on similar URLs. In the proposed system the first approach mines frequent query patterns from users search history and user profile, and the second approach identifies relevant group of queries to retrieve the exact document and extracting information from the documents for all users search.

This paper structured as follows. Section I gives introduction, Section II web information retrieval related work, section III presents mathematical model of system in section IV proposed fitness function ,in V section shows detail about proposed system, section VI presents experimental results of system and in last section conclusion is presented.

2. RELATED WORK

This section shows the concepts background pertaining to the proposed work. A major challenge for search engines is to understand users' search interest and use profile behind queries. According to that exact data retrieval and query Suggestion thus, the most important as well as challenging feature of search engine. Traditional approaches focuses on the search intent related to the query but sometimes fails to provide users' search interest related result due to ambiguity of the query that problem is solved in this project.

There are some other works to improve the results of web searching using different techniques. In [2] use famous caroler to get information from the web page using the genetic algorithm. So that information processing become fast and eventually searching become more effective. In[3]Generate the heuristic by using integrated feedback from the user and genetic algorithm so that most curate data well be taken out from web sites.[4] is another method for improving web searching this research use for information retrieval in the area of optimizing Boolean query, here Boolean logic operation for information retrieval. [5], the authors proposed a method for guiding genetic algorithms to perform information retrieval by and discuss the problem of web search and genetic algorithm application for the process of information retrieval. [6] is a the model of hybrid genetic algorithm and practical swarm optimization for web information retrieval and search optimization. In [7] paper genetic algorithm is use for clusters optimization in order to improve the query of cluster for effective personal web search, in [16] introduced query clustering approach using content words and user feedback, Combining Content and feedback similarity approach so it is efficient but it's difficult to set parameters for linear combination of two similarity metrics.

In [17] used search results for query clustering, Similarity based on ranked URL results return by search engine this approach is having better scalability, Many studies have investigated the problem of determining similar queries. For example, in [18] query alteration is developed to help users input correct words and search relevant results. The technology is widely applied by search engines online so that there is need to develop a system which extract, manage and sort that information based on user prospective that our proposed system have.

3. MATHEMATICAL MODEL

The mathematical model of the system is described as bellow Objective: provide optimize and exact view of the information from the web document according to the user query as well as profession / hobby ,also suggest query's for future search operation.

Let S be the system
such that

$S = \{I, O, F, D, Q\}$

Where I = Input
O = Output

F = Function
D = document
Q = Queries

Input: I = {QU, P, H}

Where

QU = User Query
P = User profession
H = User search history
I = Di ={d1, d2, … dn}
Di = user Data …….1< i < n

Output: O = {qk, DR, EX}

DR = {User search related Rank document}
= {DR1,DR2……DRM}
DRM = Document Ranking…….1< i < m

EX = {Extracted data from web pages}
= {E1, E2….EP}
EP = Extracted data …….1< EP < DN

QK = {Q1, Q2…….. , QK }
QK = Query suggestions…….1< i < k
FB = {F1, F2, F3, F4, F5, F6}

Function:

F1 = {Genetic algorithm with cosine similarity fitness function}

F2 = {Data extraction process}

F3 = {Automatic keyword extraction from document}

F4 = {Document filtering and ranking}

F5 = {Query suggestion}

F6 = {User profile time based operation}

4. PROPOSED SYSTEM

Flowing system architecture shows the different working modules of the proposed system for eliminating the problem of web searching process and to improve the information retrieval operation according to the user prospective hear is the new approach architecture is based on Client and server model. We are using apache-tomcat server to handle sever side operation, the system operational process goes through several steps and three module.

4.1. Module 1: Document Retrieval and Ranking

1. User registration:- To initiate the search process a user has to first register onto the system.
2. Pre-processing stage:- After user registration, user can fire search query, the initial search query goes through preprocessing stage such as ‘tokenization’, ‘stop-word’ removal, ‘normalization’, and ‘stemming’.
3. Iterative data:- The outcome of pre-processing stage is collected and ranked with respect to the search index.

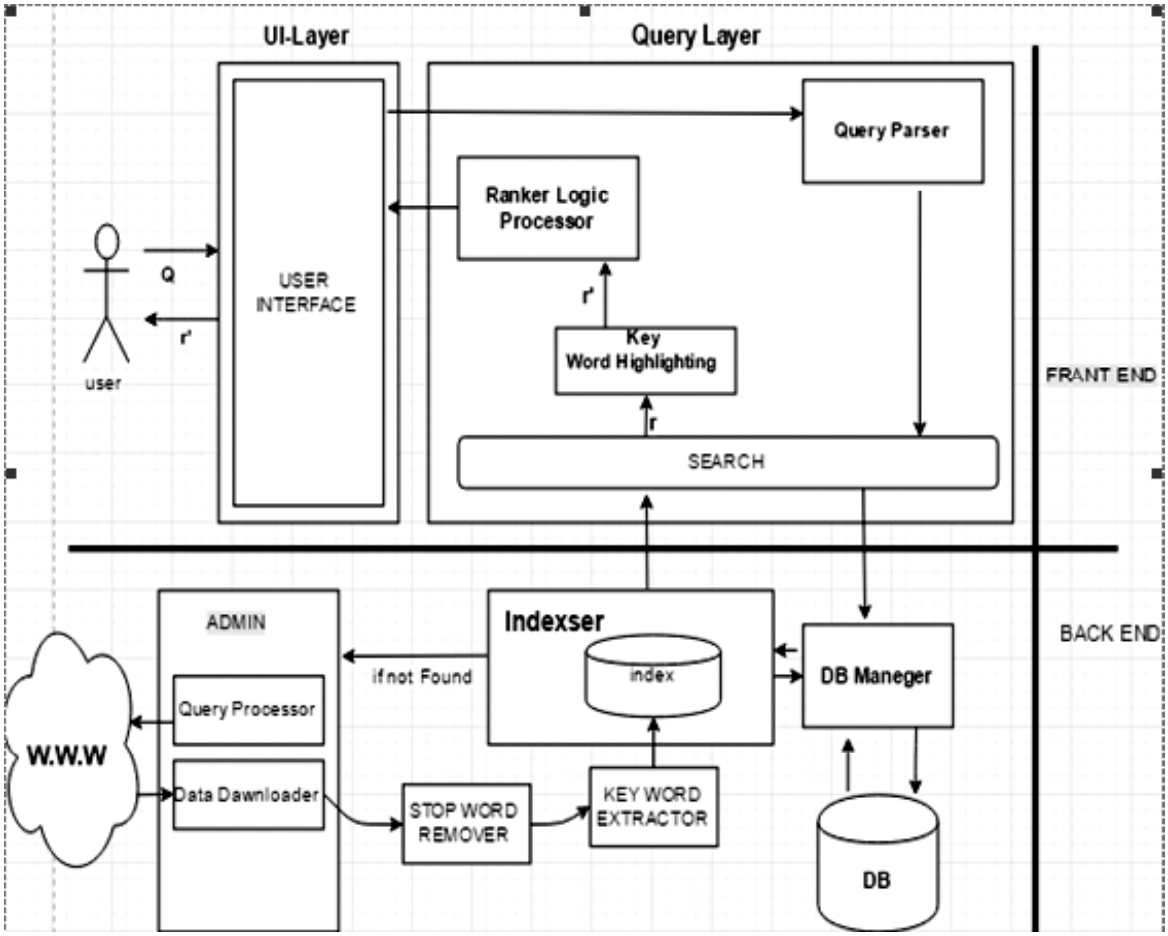

Figure 1: System Architecture

4. Genetic Algorithm:- GA will be applied on iterative ranked data. For selecting the best outcome from population selection criteria is used which is defined by cosine based similarity fitness function.
5. Threshold comparison:- For every generation of genetic algorithm, the fitness individual population is compared with threshold value.
6. Decision making:- We are using previously search data for making decision regarding relevancy of user query and collected in house data.
7. Cosine content based:- In an attempt to get more abstract outcome.

4.2. Module 2: Query Recommendation

Query Recommendation Module working is based on the parentage wise wattage of all four strategies use in our project striated bellow

1. Using User Logs:- We calculate the query loges to find user interest at the time of searching on web
2. User Profile Based:- By the help of the user profile system can understand what type of data need to serve to the user based on query.
3. User History Based:- based on user search history system can identify important part of the data that need to focus at the time of query recommendation.
4. Bi-partite Graph Based:- Relation in between queries and the URL is calculated and according to that recommended queries will get ranked.

4.3. Module 3: Data Extraction

After ranking the document by the help of the genetic algorithm in module 1, in module 3 we extract the important part of the each document effectively separately to save the user operational time by finding cosine similarity term based ratio in the document.

5. PROPOSED FITNESS FUNCTION

A fitness function is a special type of objective function that is use for summarization, as a single figure of merit, how close a given design solution is to achieving the set aims.

In particular, in the fields of genetic algorithms and genetic programming, each specific design solution is mostly represented as a string of numbers (generally represented as a chromosome). After each round of simulation or testing, the idea is to delete the 'n' worst design solutions, and to breed 'n' new ones from the best design solutions. Each design solution, therefore, needs to be awarded a figure of merit, to indicate how close it came to meeting the overall specification, and this is generated by applying the fitness function to the test, or simulation, results obtained from that solution.

The cosine similarity calculation formula is use as a fitness function is declared as follows:

$$\cos\theta = \frac{|A \cap B|}{|A|^{1/2} . |B|^{1/2}} = \frac{\sum_{i=1}^{n} Ai.Bi}{\sqrt{\sum_{i=1}^{n} (Ai)^2} . \sqrt{\sum_{i=1}^{n} (Bi)^2}}$$

where A is the document vector and B is the query vector.

6. EXPERIMENTAL SETUP AND RESULT

We use genetic algorithm for primarily retrieved document filtering with respect to the user queries, for that cosine similarity fitness function is used, also by user search history we are able to find user interest to handle ambiguity of the searching as the next step we perform query suggestion operation by the help of user search loges, search history, profile and queries click graph by adding a number of search logs. That combination of the different methodology stated above reduce the effect of outliers and noise, also improves the performance of the system.

The proposed system platform and technology used is as follows:

- Base Operating System: Windows 7
- Databases: My SQL
- Web Server: Apache
- Language: Java
- Browser: IE8 & above, Mozilla Firefox, Google Chrome, Opera etc.

At the initial state of the experiment we compare different similarity methods such as cosine similarity, jaccard and dice similarity with our approach of document filtering according to the different categories of the user profile and query's such as business, agricultural, computer, social, the result potent is shown in the following table.

According to the above table result graphical notation is obtained as bellow.

According to the above graphical notation and result table shows that our approach works far better than the other similarity algorithm

Table 1
Comparison table

Query/Category	Cosine Similarity	Dice Similarity	Jaccard Similarity	Genetic Algorithm		Genetic Algorithm with cosine Similarity approach	% increase in relevancy
				Crossover Probability	Mutation Probability		
business	0.0958927	0.0371353	0.00612245	0.1	0.01	0.21353961767801954	10.480629844098607
software	0.285801	0.035533	0.00686499	0.1	0.01	0.07573096110550687	3.793553344922079
company	0.0958927	0.0319149	0.00510204	0.1	0.01	0.04787310224906509	2.282356649575779
marketing	0.0101331	0.0292505	0.00472335	0.2	0.02	0.0712280831356229	3.495841727136168
anna hazare anti corruption	0.0587696	0.0653061	0.0109785	0.1	0.01	1.1346334569000805E-4	0.005512905464437466
Data Structure	0.0135559	0.0159314	0.00178398	0.2	0.02	0.09149520077652332	3.7860917037601793
quality	0.351755	0.0305344	0.00572082	0.1	0.01	0.06126684307333077	2.5352352781631637

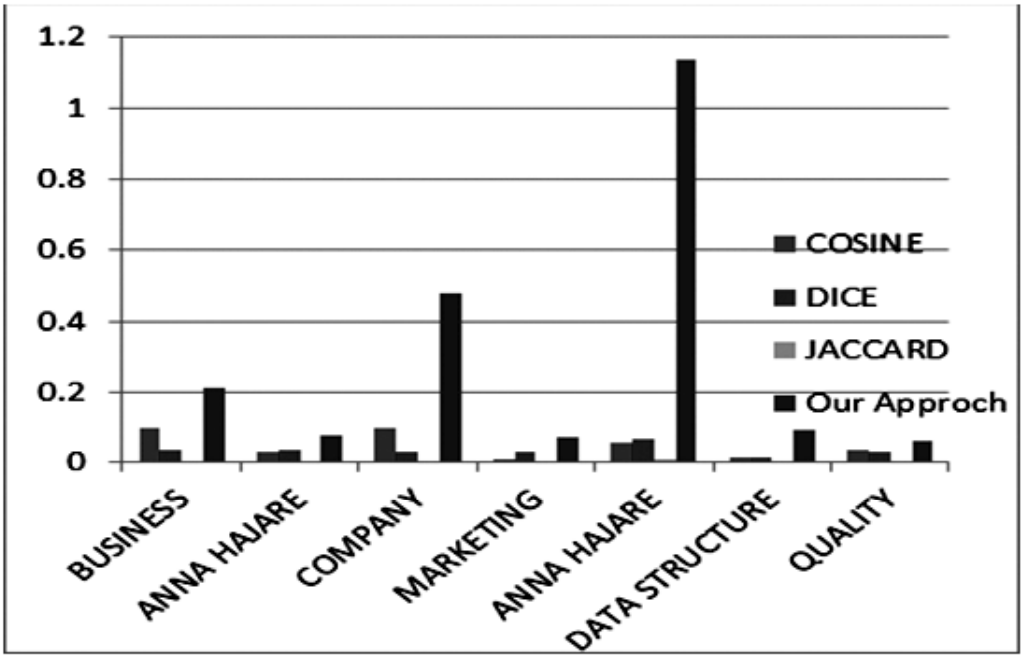

Figure 2: Result Graph

7. CONCLUSION

- According to the result we found that proposed system gives best result for the particular user query.
- Fitness criteria based on cosine similarity gives the best performance of the genetic algorithm for optimizing the searching process.