GA BASED SEARCH OPTIMIZATION FOR INFORMATION RETRIEVAL

[1]ANIKET M. AKARTE, [2]PRADNYA V. KULKARNI

[1,2]Dept. of computer engineering, MAEER'S MIT Pune, India
E-mail: [1]a1.aniket@hotmail.com, [2]pradnya.kulkarni@mitpune.edu.in

Abstract— As continues growth in web data the complexity and challenges in exact data extraction is increase, to address that issue this paper presents a combine approach of Genetic Algorithm and cosine similarity for optimizing the process of Web Information Retrieval and personalizing the user search by using deduce sensing mechanism so that quality and the accuracy of information collected is improved.

Keywords— Genetic Algorithm ; Cosine Similarity ; Web Search ; Information Retrieval.

I. INTRODUCTION

The most promising information source in the world, the World Wide Web is still expanding rapidly.[1] The capacity of storage device is increase and cost is decrease there is tremendous growth in database of all sorts. [2]This explosive growth has led to huge, fragmented and become easy to collect and store information in document collection; it has become increasingly difficult to retrieve relevant information from this large document collection, the search engines play a very important role during this process.

Search engines aims to process the enormous information in some collection of document then create an index for quick search. Basically, the index is an inverted file that maps each word in the collection to the set of documents containing that word. Web Information Retrieval is to better cater to the information need of the user. Personalization of Web Search is used for intelligent web search based on knowledge obtained by mining the web data. This paper presents new approach of information retrieval based on two technique known as genetic algorithm and cosine similarity to optimize user search as well as in this research we use new technique known as deduce sensing mechanism whose improve the accuracy of searching according to the user perspective.

Genetic algorithm is a type of the probabilistic search algorithm which mimic the process of natural selection of sepsis of organism known as Darwinian theory deal with very large search spaces as in information retrieval the genetic algorithm start from a new initial population for each query and select best individual obtained after giving number of generation in the end of performance of the experiment is computed as the average of the result portend by the best individual of each query. Genetic algorithm able to significantly improve the performance in reasonable amount of time so that genetic algorithm can perform more effective search.

II. RELATED WORKS

There are some other works to improve the results of 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] in this paper genetic algorithm is use for clusters optimization in order to improve the query of cluster for effective personal web search, so that search process become more effective as compared to that our proposed system have several advantages stated as bellow.

The advantages of the proposed method are as follows:

- Genetic algorithm with cosine similarity fitness function improve the relevancy and quality of the retrieved document effectively as compared to the classical method of the information retrieval.
- Genetic algorithm are robust type of search optimization algorithm so that it will not get trapped in local maxima and produce globally optimize result as per the user search.
- As we uses deduce sensing mechanism for future searching and query formation so that user can navigate and search more accurately

as compared to the classic web searching process.

III. PROBLEM STATEMENT

As comparing to the previous method proposed combine approach help to extract exact quality data according to the user search, cosine similarity fitness function improve the performance of genetic algorithm in information retrieval as well as by the help of deduce sensing mechanism the user can classify the category of data that he want to search and reduce the ambiguity in searching so the overall processing of information retrieval is improved by new approach stated in this paper.

IV. PROPOSED SYSTEM

Figure 1 shows the methodology of the system. The user enters the query on the search engine. 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 run the server process, the filtration process goes through several steps.

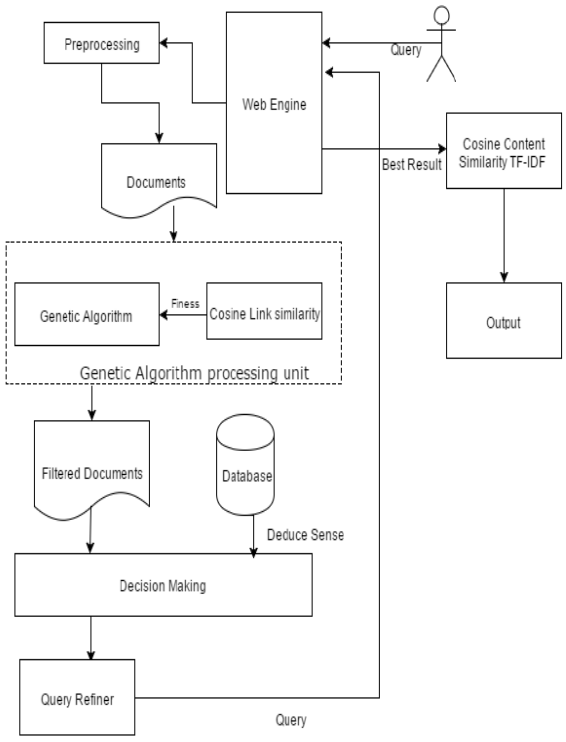


Figure 1: Architecture of Proposed System

Step 1. User registration: To initiate the search process a user have to first register onto the system.

Step 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'.

Step 3. Iterative data: The outcome of pre-processing stage is collected and ranked with respect to the search index.

Step 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.

Step 5. Threshold comparison: For every generation of genetic algorithm, the fitness individual population is compared with threshold value.

Step 6. Decision making: We are using previously search data for making decision regarding relevancy of user query and collected in house data.

Step 7. Cosine content based: In an attempt to get more abstract outcome, Cosine content based similarity will be used on the outcome of previous step.

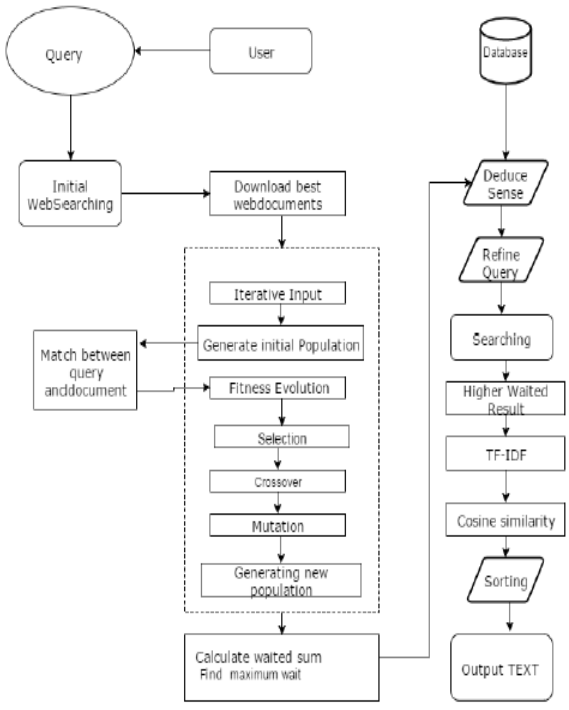


Figure 2: System Flow

Above diagram shows the detail working of the proposed system with respect to the different stapes stated as above. That helps to get exact data according to the user search.

V. ALGORITHM

Here we are use the two basic method for overall system processing that are genetic algorithm and cosine similarity from that we again use the two basic

forms of cosine similarity that is cosine link similarity for training the fitness function and cosine content based similarity for document level similarity measurement due to that overall operation of the system is improved. We use the concept of the deduce sensing mechanism so the accuracy of the searching process is improved so we gate most accurate result as compared to the normal searching through the website ,genetic algorithm is basically use in the information retrieval system for search optimization by the fitness criteria set by the cosine link similarity.

Table 1 shows whole of the above procedure summarized in the following algorithm:

Table 1. Algorithm for GA-CS

Input:
Query entered by the user.
User's registered details.
Process:
Preprocess the query entered by the user.
Generate all the possible combinations.
Individuals are created from all the combinations.
Perform following to find maximum weight individual by Applying Genetic Algorithm.
Generate initial population by passing the individuals to GA.
Do for one generation.
Calculate fitness function based on cosine link similarity mechanism for each individual in the population.
Apply Tournament selection operator for choosing winners to fight by calculating each individual's fitness.
Let two and two fight.
The winners make children after performing crossover.
Children make new population.
Randomly select an individual for mutation.
Mutate bit from 0 to 1 or 1 to 0 by selecting random position of the individual.
Final population is created after including the mutated individual.
Calculate weighted sum of each individual in the population.
Find maximum weight individual.
If maximum weight is greater than the threshold value deduces first sense from the database else deduce second sense.
Apply cosine similarity content similarity over document.
Output:
Refined query, Required text data

VI. PROPOSED FITNESS FUNCTION

Fitness function acts as a selection operator for genetic algorithm according to the fitness value the genetic algorithm evaluates good solution for processing, in this research we use two different fitness functions for two different stage of genetic algorithm operation.

(a) for first part fitness function in genetic algorithm use measure of cosine similarity between query supplied by the user and the chromosome value as per that fitness value is calculated by the formula

$$\frac{\sum_{t=1}^{t} X_i.Y_i}{\sqrt{\sum_{i=1}^{t} X_i^2 . \sum_{i=1}^{t} Y_i^2}}$$

where Xi represents the weight i in the chromosome ,Yi is real representation of the weight of term of the query where t is total number of term and query vector as in given chromosome through the calculating value of similarity measure always lies between 0 and 1 according to cosine similarity and the query.

(b) for second part of genetic algorithm uses a new Fitness criteria represents for similarity calculation by

$$\sum_{i=1}^{t} X_i - Y_i$$

That states the difference between the term values of given chromosome and query supplied by the user

VII. EXPERIMENTAL SETUP AND RESULT DISCUSSION

In this research we tried to apply combine approach of genetic algorithm and cosine similarity for improving the processing of information retrieval from the web pages we also deal with ambiguous query processing using deduces sense mechanism this proposed system has been develop on java platform figure 1 shows the diagrammatic representation of the system fitness value of relevancy between two document

Hear we use training sample is compare document in category A and two documents in other categories

GA Operation- genetic algorithms use three different operators such as selection, crossover, and mutation

• Selection - as i stated earlier in this research fitness value based on cosine similarity from that fittest chromosome value are to be selected and duplicated for the next generation.

• Crossover Operation- generally cross over operation produce fittest offspring value by combining two selected parents which acts as a new generation.

Parent Chromosome I: 11112011011 (10 words)
Parent Chromosome II: 111011210 (8 words)
Randomly selected functional value for 1 and randomly selected functional value 2:
- cross point (1).1:3 - cross point (1).2:5
- cross point (2).1: 4 - cross point (2).2: 8

Parent Chromosome 1: 11112012011
(Exchanging genes in boldface)
Parent Chromosome 2: 111012210

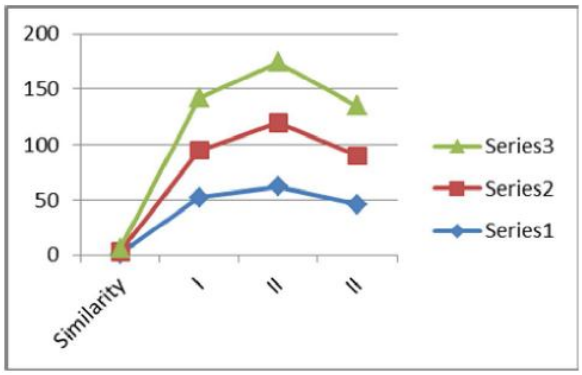
Crossover:
Offspring 1: 1111221012011 (13 words)
Offspring II: 1110120 (7 words)
Training documents according to given
Process:
String Pairs (word, occurrence)
Doc. 1 100100011211 (0, 4) (1, 7) (2, 1) category A
Doc. 2 11111201201 (0, 2) (1, 7) (2, 2) category A
Doc. 3 110010112201 (0, 4) (1, 6) (2, 2) category A
Doc. 4 00110001100 (0,8) (1, 4) other category

Doc. 5 101010101010 (0, 6) (2, 6) other category
Population of chromosomes in category A:

String	Pairs (word, occurrence)
Chromosome I 1000101001	(0, 6) (1, 4)
Chromosome II 11112012011	(0, 2) (1, 7) (2, 2)
Chromosome III 111012210	(0, 2) (1, 5) (2, 2)

Similarity	1	2	3	4	5
I	52	43	47	63	66
II	62	58	55	37	53
II	46	44	45	32	43

Graphical representation of similarity value is shown in following



So the genetic algorithm calculated the cosine similarity fitness criteria with the chromosome notation are stated as bellow.
Fitness (I) = 46.018 Fitness (II) = 58.201 Fitness (III) = 44.019

CONCLUSION

- From result we observe that when cosine similarity is used to calculate fitness score in Genetic algorithm, improves the accuracy of retrieved documents.
- By compares different genetic algorithm strategies with different similarity measure we fund that cosine similarity gives best result among all similarity measure.
- The process of query training from database improves efficiency of searching and helps in reformulating the query to get more exact result.
- Combine approach improve the efficiency and accuracy of information retrieval process.

ACKNOWLEDGMENT

I am extremely thankful to Dr. V. Y. Kulkarni Head, Department of Computer Engineering for valuable guidance,
Constant encouragement and inspiration during each and every step of my project work throughout the semester. I owe a great deal of love to my family for their blessings and consistent moral support.

REFERENCES

[1] David C. Anastasiu and George Karypis "L2AP: Fast Cosine Similarity Search with Prex L-2 Norm Bounds"30th IEEE International Conference on Data Engineering, Chicago, IL,(ICDE)2014,DOI:10.1109/ICDE.2014.6816700,pp 784-759.

[2] Swe Swe Nyein, "Mining Contents in Web Page Using Cosine Similarity",Computer Research and Development (ICCRD), 2011 3rd International Conference on (Volume:2) Shanghai, DOI:10.1109/ ICCRD.2011.5764177, IEEE 2011, pp 472-475.

[3] Anu Kundu , Sona Malhotra,"Information Retrieval using Web",Journal of Global Research in Computer Science (IGRCS),Volume 2, No. 4, April 2011,ISSN: 2229-371X,pp 1-4.

[4] Manish Sharma , Mr. Rahul Patel "Applying Genetic Algorithm in Text to Matrix Generator" International Journal of Computer Science and Information Technologies(IJCSIT),ISSN 0975-9646, Vol. 5 (1) , 2014, pp 32-34 .

[5] Sapna Chauhan, Pridhi Arora ,Pawan Bhadana "Algorithm for Semantic Based Similarity Measure" International Journal of Engineering Science Invention(IJESI),ISSN:2319-6734,Volume 2 Issue 6 June. 2013,pp 75-78.

[6] Manoj Chahal, Jaswinder Singh "Eective Information Retrieval Using Similarity Function: Horng and Yeh Coefficient", International Journal of Advanced Research in Computer Science and Software Engineering(IJACSSE) Volume 3, Issue 8, August 2013,ISSN: 2277-128X,pp 1-6.

[7] Marina Litvak,Mark Last A,Menahem Friedman "A new Approach to Improving Multilingual Summarization using a Genetic Algorithm", ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Sweden,ACM 2010,pp 927-936.

[8] K.Deepa, M. Senthamil Selvi, R.Rangarajan Automatic Threshold Selection using PSO for GA based Duplicate Record DetectionInternational Journal of Computer Applications(IJCA),Volume 62 Number 4,DOI: 10.5120/10068-4674,ISSN:0975 8887,pp 22-27.

[9] Wafa. Maitah, Mamoun. Al-Rababaa and Ghasan. Kannan "Improving The Eectiveness Of Information Retrieval System Using Adaptive Genetic Algorithm"International Journal of Computer Science and Information Technology (IJCSIT) Vol 5, No 5, October 2013,DOI: 10.5121,pp 271-280.

[10] Roopak.S, Tony Thomas "A Novel Phishing Page Detection Mechanism Using HTML Source Code Comparison and Cosine Similarity"Advances in Computing and Communications (ICACC),2014 Fourth International Conference IEEE 2014, Cochin, DOI: 10.1109/ICACC.2014.47,pp 167-170.

[11] Digvijay B. Gautam, Pradnya V. Kulkarni "Cosine Similarity Measure and Genetic Algorithm for extracting main content from web documents",International Journal on Advanced Computer Theory and Engineering (IJACTE),ISSN : 2319-2526, Volume-3, Issue -6, 2014,pp 1-5.

[12] J. Usharani, Dr K Iyakutti, "A Genetic Algorithm based on Cosine Similarity for Relevant Document Retrieval" International Journal of Engineering Research and Technology(IJERT), Vol.2 - Issue 2 (February - 2013),ISSN: 2278-0181,pp 1-5.

[13] Vikas Thada and Dr Vivek Jaglan "Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm" International Journal of Innovations in Engineering and Technology (IJIET),SSN: 2319-1058,pp 1-4.

[14] Firas Alabsi and Reyadh Naoum "Fitness Function in Genetic Algorithm based Information Filtering - A Survey",International Journal of Computer Science and Mobile Computing (ICMIC13), December- 2013,ISSN 2320088X, pp 80-86.

[15] Wafa. Maitah, Mamoun. Al-Rababaa and Ghasan. Kannan "Improving The Effectiveness Of Information Retrieval System Using Adaptive Genetic Algorithm" International Journal of Computer Science and Information Technology (IJCSIT), Vol 5, No 5, October 2013,DOI : 10.5121/ijcsit.2013.5506,pp 1-15.

[16] Alfred V. Aho and Margaret J. Corasick Bell Laboratories, "Efficient String Matching An Aid to Bibliographic Search", communications of the ACM, Vol. 18 No.6, June1975,New York, NY, USA, DOI:10.1145/360825.360855,pp 333-340.

[17] Amit Chandel, P.C.Nagesh, Suita Sarawagi, "Efficient Batch Top-k for Dictionary-based Entity Recognition", 22nd International Conference Data Engineering(ICDE)2006 IEEE,DOI:10.1109/ICDE.2006.55,pp 1-28.

[18] Amit Singhal, "Modern Information Retrieval: A Brief Overview", IEEE Computer Society Technical Committee on Data Engineering,Vol. 24, No. 4. (2001),DOI:10.145/361219.36122001, pp. 35-42.

★★★