



Fusion of Similarity Coefficients for Web Information Retrieval Using Genetic Algorithm

Dr. Vikas Thada, Aman Jatain, Utpal Srivastva
Assistant Professor, Amity University
Gurgaon, Haryana, India.

Abstract: Crawling is a process in which web search engines collect data from the web. Focused crawling is a special type of crawling process where crawler look for information related to a predefined topic. In this paper a method for finding out the most relevant document among a set of documents for the given set of keyword is presented. Relevance checking is done with the help of Rogers-Tanimoto, MountFord and Baroni-Urbani/Buser similarity coefficients. The weights obtained using these coefficients were then analyzed as to which 2 or 3 coefficients can be used for fusion. In the process of fusion results obtained from several similarity coefficients are combined using one of the fusion rules. Out of the three popularly used fusion algorithms: sum, min and max, the research work used sum and max. The method uses genetic algorithm to show that the average similarity of documents to the query increases when Probability of mutation is taken as low and Probability of crossover is taken as high. The method does the performance analysis of different similarity coefficients on the same set of documents and applies ranking to the documents whose relevancy is highest among the three coefficients by applying fusion algorithm on them.

Keywords: Crawl, Coefficient, Engine, Fusion, Genetic Algorithm, Growth, Similarity, Retrieval

I. INTRODUCTION

Design of most focused crawlers is based on the vector space model. The model is used to judge the evenness of web pages and general web search algorithms. The relevance in turn works as guide in following target links (Wikipedia, 2014). One of the most important modules of search engine is ranking module. The task of ranking module is to assign some ranking score to relevant pages using some criterion. Output of ranking module is an ordered set of pages according to their rank i.e. pages with high rank are near the top of the list and low rank pages are at the bottom of the list. These pages are then presented to the user in their ranking order. A GA based approach using Rogers-Tanimoto, MountFord and Baroni-Urbani/Buser similarity coefficients is taken in this paper for ranking the retrieved documents.

II. GENETIC ALGORITHM

GAs are search algorithms that follow the concept of natural selection and genetics [3]. GA are powerful and very efficient search and optimization techniques motivated by the natural selection theory of Darwin [4]. Genetic Algorithms [5] are based on the principle of heredity and evolution which claims “in each generation the stronger individual survives and the weaker dies”. Therefore, each new generation would contain stronger (fitter) individuals in contrast to its ancestors. The input to the process of GA’s number of possible solutions known as population which is of constant size. In each generation/iteration performance of each chromosome in current population is calculated (depends upon problem in contest) known as fitness and new population evolves. In the reproduction phase the process of GAs (fitness evaluation, selection, crossover, and mutation) continues with chromosomes of higher fitness values resulting in new population. Chromosomes with lower fitness values are discarded. Process of GA is again applied to this population and this continues until we get an optimal solution for the given problem [6].

For getting an optimal value among many different solutions fitness function is used. The function is written depending upon problem in consideration to judge the fitness of a solution i.e. how optimal it is among other possible solutions. The problem of IRS using GA is to retrieve documents using this fitness function. For finding the relevant document on the basis of some similarity measures we can have number of relevancy methods,

Table 1 Coefficients Used As Fitness Function In Research

S.N	Coefficient Name	Similarity Formula
1.	Rogers & Tanimoto	$(p+s) / (p+2*(q+r)+s)$
2.	Baroni-Urbani/Buser	$(p+\sqrt{p*s}) / (p+q+r+\sqrt{p*s})$
3.	Mountford	$2*p/(2*q*r+p*q+p*r)$

For the calculation of similarity metric we have defined few parameters p,q,r and s as ($n = p+q+r+s$) as shown in table 2.

p= (x=1 and y=1) (total match)
 q= (x=1 and y=0) (single match)
 r= (x=0 and y=1) (single match)
 s= (x=0 and y=0) (no match)

Table 2 Variables Used to Calculate Binary

	y=1	y=0
x=1	p=1/1 in both A and B	q=1/0 only in A
x=0	r=0/1 only in B	s=0/0 in none of A and B

In table 2, where A and B may be any query or document represented in binary form.

III. RESEARCH TOOLS

This section is about the introduction of tools which helped us to find the top keywords from relevant document. These keywords actually used for making the chromosomes; backbone of GA in which the research implementation work was carried out. The research work has used Text Analyzer and Keyword Density Checker as research tools . Both the tools were used to find out top most keywords present on webpage with highest frequency.

IV. EXPERIMENTAL SETUP & RESULTS

In general the web documents are encoded in strings of 0's and 1's as shown in the figure 4 The documents have been obtained using some search query. The set of documents also contains some non-relevant documents too. In the experimental setup of GA 50 documents have been retrieved for each different query. This set of 50 documents serves as document database. The query is also encoded in terms of string of 0 and 1. Process of experiment is as follows [15].

- Query is input to the Google search engine.
- Top frequent keywords of each retrieved document are extracted using Textalyser or Keyword Density Checker online tool and making list of keywords.
- Generate initial population by encoding retrieved documents to chromosomes for each query.
- Initial population now consists of these encoded documents. Pass this initial population to GA algorithm.
- Either some maximum generation or some predefined fitness value is not achieved repeat step 4. At the end of this step we get an optimized document for retrieval.
- Optimized document chromosome can be decoded and document can be retrieved from stored documents related to each query.
- Apply fusion MAX and SUM algorithm to results obtained using similarity coefficients.
- Rank the documents on the basis of fitness value as obtained in step vii.

A. Experimentation

The research work conducted the tests for 15 different queries and 30 documents for each query. The experimentation tests for all 15 queries with RT, MF and BUB coefficients chosen as fitness function. A complete MATLAB code has been written with roulette wheel selection operator, random point crossover with different rates of crossover and mutation. Experiment conducted with various GA parameters as: probability of crossover ProC = 0.7, 0.8 and 0.9 along with probability of mutation (ProM =0.01, 0.10, and 0.30) to do performance analysis of this GA based retrieval system. The efficiency parameter is average relevance. Average relevance from three fitness functions are compared and best combination for crossover and mutation probability for any one among three similarity function is selected as best . Fusion algorithm have been applied onto this results to get combine ranking. Using this best selected similarity function the document's rank before and after the algorithm are checked and new ranks to document are assigned. The retrieval process now retrieves documents using this new ranking scheme.

Table 3: Queries Used in Research Work

S.N	Chromosome Length	Keywords
1.	33	Delhi, Gang, Rape, Case
2.	39	Bomb, Blast, Boston, Marathon
3.	37	Search ,Engine, Optimization,Seo

4.	49	Anna, Hazare, Anti, Corruption
5.	53	Osama, Bin, Laden, Death, Killed
6.	47	Stock, Market, Mutual, Fund
7.	55	Fiber, Optic, Information, Technology
8.	66	Health, Medicine, Medical, Disease
9.	44	Artificial, Intelligence, Neural, Network
10.	40	Rajasthan, Royals, Amity, University
11.	49	Remote, Method, Invocation, RMI
12.	61	Gang, Rape, Kangaroo, Court
13.	69	Aam, Aadmi, Party, Arvind, Kejriwal
14.	50	Genetic, Algorithms, Optimization, Techniques
15.	64	Search, Engine, Web, Crawler

B. Results

In this section we show the results in tabular and graphical form for all three similarity coefficients viz. RT, MF and BUB. We vary probability of mutation ProM from 0.01 to 0.10 and 0.30 and keep ProC=0.5, 0.7 and 0.9 for all the values of ProM. Keeping together ProC and ProM we get total nine possible combinations for checking the performance of said similarity coefficients on retrieved documents. Out of nine possible results we got best was for ProC=0.9 and ProM=0.1. Tables and figures for rest eight combinations have not been shown.

Further it was observed that best average maximum fitness is given by Rogers-Tanimoto coefficient. Figure plotted using the above table is shown in Figure1.

C. Fusion of Similarity Coefficients

Numbers of thoughts for fusion came out from the idea of fusing different mutually exclusive web information retrieval (Web-IR) system components to retrieve and improve the efficiency and retrieval performance against best individual component. It is always worthy of fusing many similarity coefficients together as different measures may make use of different sources of evidence. Web-IR system considers several cases of fusing similarity coefficients with different presentation of data and consider the reasons for positive or negative results in terms of the similarity distributions [17].

It was found that fusion using SUM and MAX generally gives a higher number of ranking to individual web documents against a query as compared to the individual similarity measures. It is also proved that the SUM fusion is significantly better and followed closely by MAX. For each query, the rankings produced by fusing the 3-similarity coefficients using MAX and SUM fusion algorithms were averaged to give a new ranking. Based on this new ranking, it was found that the best performing coefficient combinations returned more relevant results compared to the best individual coefficient.

Table 4 Fusion of RT, MF and BUB using MAX algorithm

QUERY	(RT,MF	RT,BUB	MF,BUB	RT,MF,BUB
Query1	0.6309	0.6309	0.6141	0.6309
Query2	0.5442	0.5446	0.5446	0.5446
Query3	0.5729	0.5729	0.5462	0.5729
Query4	0.5541	0.5541	0.4695	0.5541
Query5	0.5381	0.5381	0.4509	0.5381
Query6	0.5348	0.5348	0.4702	0.5348
Query7	0.5462	0.5462	0.4507	0.5462
Query8	0.5268	0.5268	0.4113	0.5268
Query9	0.5852	0.5852	0.5198	0.5852
Query10	0.5838	0.5838	0.5171	0.5838
Query11	0.5604	0.5604	0.4397	0.5604
Query12	0.5294	0.5294	0.4641	0.5294
Query13	0.5345	0.5345	0.4241	0.5345
Query14	0.5552	0.5552	0.4808	0.5552
Query15	0.4987	0.4987	0.4518	0.4987

Table 5 Fusion of RT, MF and BUF using SUM algorithm

Query	RT,MF	RT,BUB	MF,BUB	RT,MF,BUB
Query1	0.46285	0.6225	0.45445	0.5133
Query2	0.38745	0.5444	0.38765	0.4398
Query3	0.38	0.55955	0.36665	0.4354
Query4	0.3522	0.5118	0.3099	0.3913
Query5	0.33195	0.4945	0.28835	0.3716
Query6	0.33425	0.5025	0.30195	0.3796
Query7	0.3322	0.49845	0.28445	0.3717
Query8	0.3074	0.46905	0.24965	0.342
Query9	0.3759	0.5525	0.3432	0.4239
Query10	0.39195	0.55045	0.3586	0.4337
Query11	0.348	0.50005	0.28765	0.3786
Query12	0.3071	0.49675	0.27445	0.3594
Query13	0.308	0.4793	0.2528	0.3467
Query14	0.35045	0.518	0.31325	0.3939
Query15	0.29335	0.47525	0.2699	0.3462

D. Ranking of Documents

To rank the documents using GA on the basis of fitness obtained using RT for the above parameters of GA for all queries, we run the MATLAB code for a fixed number of generations for RT and obtain the fitness values for each documents. Fusion operation were applied on the data shown in table 4. The result thus obtained are shown in table 5 and 6. By looking at the table 5 and table 6 it is clearly visible that MAX function when applied to all three similarity coefficients gives best result. Further SUM fusion algorithm applied to RT and BUB similarity gives best fitness values. The documents are ranked on the basis of this obtained fitness values. We show this only for first query Q1={ Delhi, Gang, Rape, Case} in table . The hyperlinks for each of the documents are not shown in fact we name them as Di (i=1 to 30).

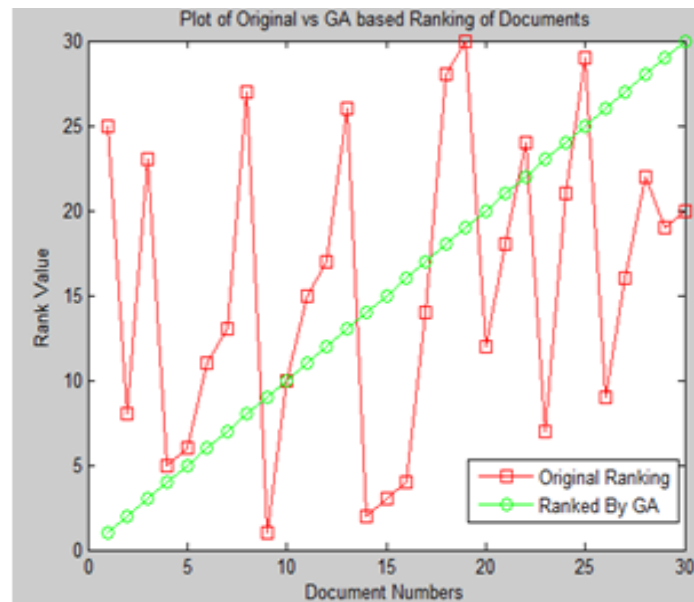


Figure 1 Comparative ranking of Original and GA based using MAX fusion algorithm.

The comparison of old and new rankings and percentage improvement is shown graphically in the figures below: The figure 5 and figure 6 shows that document having rank 25 has got new rank 1 and percentage improvement in rank is 80. Similarly another document having old rank as 8 has moved to new rank 2 and got 20 % increment in rank because of our applied GA with RT.

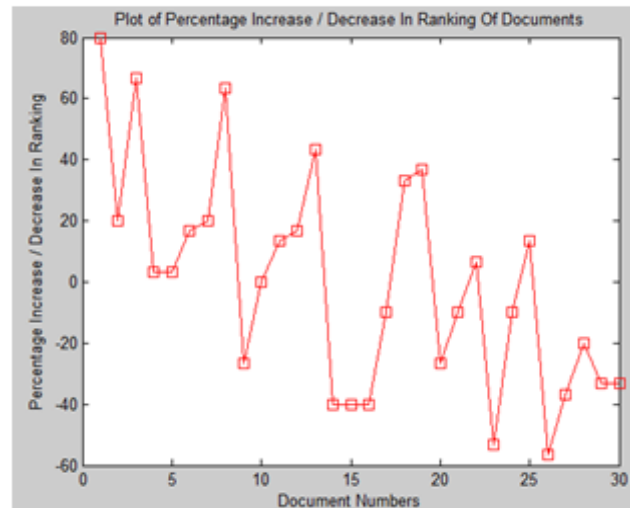


Figure 2 percentage increase/decrease in ranking of documents

We have also checked online by opening pages with new rank and they were providing the good result as required by the corresponding query. Similar type of work was carried out for other queries and their related documents but as not to extend the length of this research paper we have provided ranking table and figure only for one query.

V. CONCLUSION AND FUTURE WORK

The research work has calculated average maximum fitness value using RT, BUB and MF similarity coefficients. In the work ProC and ProM was varied from 0.7-0.9 and {0.01, 0.10, 0.30} respectively. It has been shown that average relevancy of retrieved document increases when PC is increased and ProM is decreased. Using the best average fitness value among RT, MF and BUB the research work applies new ranking to retrieved documents. The goal was to design the web ranking module using genetic algorithm and focused crawling and good results are a manifest to this research work. In this way searching has been made easy, so more relevant document or web pages can be retrieved easily. Results show that average relevance of document increased up to 63%. In this way if focused crawler have key set with more relevancy then the retrieved data is more relevant for local collection of a search engine which improves the crawling performance.

REFERENCES

- [1]. Novak B, "A Survey Of Focused Web Crawling Algorithms Proceedings of SIKDD", pp. 55-58, 12-15 Oct 2004.
- [2]. http://www.wikipedia.org/Web_Crawler
- [3]. B.Klabbankoh, O.Pinngern. "Applied genetic algorithms in information retrieval" Proceeding of IEEE, pp.702-711, Nov 2004
- [4]. S.S.Satya and P.Simon, "Review on Applicability of Genetic Algorithm to Web Search," International Journal of Computer Theory and Engineering, vol. 1, no. 4, pp. 450-455, 2009.
- [5]. Shokouhi, M.; Chubak, P.; Raeesy, Z. "Enhancing focused crawling with genetic algorithms", Vol: 4-6, pp.503- 08, 2005.
- [6]. V.Thada, S.Dhaka,"Genetic Algorithm based Approach to Solve Non Fractional (0/1) Knapsack Optimization Problem", International Journal of Computer Applications, vol 100 – No.15, Aug 2014.
- [7]. www.sequentix.de/gelquest/help/distance_measures.htm.
- [8]. V.Consonni and R. Todeschini, "New Similarity Coefficients for Binary Data", Communications in Mathematical and in Computer Chemistry, pp.581-592, 2012
- [9]. H.Wolda, "Similarity Indices, Sample Size and Diversity", Oecologia-Springer-Verlag, pp. 296-302, 1981
- [10]. M.A.Kauser, M. Nasar, S.K.Singh, "A Detailed Study on Information Retrieval using Genetic Algorithm", Journal of Industrial and Intelligent Information vol. 1, no. 3, pp.122-127 Sep 2013.
- [11]. http://en.wikipedia/wiki/Fitness_Proportionate_Selection
- [12]. J.R. Koza, "Survey of Genetic Algorithms and Genetic Programming", Proceedings of the Wescon, pp.589-595, 1995
- [13]. <http://textalyser.net/>
- [14]. <http://www.webconfs.com/keyword-density-checker.php>.
- [15]. V.Thada, V.Jaglan, "Use of Genetic Algorithm in Web Information Retrieval", International Journal of Emerging Technologies in Computational and Applied Sciences, vol.7,no.3,pp.278-281, Feb,2014
- [16]. Y.A.A. Ali, "Fusion of Similarity Measures Using Genetic Algorithm for Searching Chemical Database", Faculty of Computer Science and Information System University Technology Malaysia, pp.8-9, Nov 2007.