A GENETIC ALGORITHM APPROACH FOR IMPROVING THE AVERAGE RELEVANCY OF RETRIEVED DOCUMENTS USING JACCARD SIMILARITY COEFFICIENT

Mr. Vikas Thada*

Mr. Sandeep Joshi**

ABSTRACT

The rapid growth of the world-wide web poses unprecedented scaling challenges for general-purpose crawlers and search engines. A focused crawler aims at selectively seek out pages that are relevant to a pre-defined set of topics. Besides specifying topics by some keywords, it is customary also to use some exemplary documents to compute the similarity of a given web document to the topic. In this paper we present a method for finding out the most relevant document for the given set of keyword by using the method of similarity measure of jaccard coefficient. We use genetic algorithm to show that the average similarity of documents to the query increases when initial set of keyword is expanded by some new keywords. The similarity coefficient for a set of documents retrieved for a given query from google are find out then average relevancy is calculated. The query is expanded with new keywords and we show that average relevancy for this new set of documents increases.

Keywords : *algorithm ,average, coefficient, genetic, jaccard,, relevancy.*

*M.Tech(CS) IV Sem , Department of Computer Engineering, Sobhasaria Engineering College, Sikar(Rajasthan).

** HOD(CS &IT), Department of Computer Engineering, Sobhasaria Engineering College, Sikar(Rajasthan).

1. INTRODUCTION

The rapid growth of the World-Wide Web poses unprecedented scaling challenges for general-purpose crawlers and search engines. The first generation of crawlers on which most of the web search engines are based rely heavily on traditional graph algorithms, such as breadth-first or depth-first traversal, to index the web. A core set of URLs are used as a seed set, and the algorithm recursively follows hyperlinks down to other documents. Document content is paid little heed, since the ultimate goal of the crawl is to cover the whole Web [1]. The motivation for focused crawler comes from the poor performance of general-purpose search engines, which depend on the results of generic Web crawlers. So, focused crawler aim to search and retrieve only the subset of the world-wide web that pertains to a specific topic of relevance. Focused crawling search algorithm is a key technology of focused crawler which directly affects the search quality.

The ideal focused crawler retrieves the maximal set of relevant pages while simultaneously traversing the minimal number of irrelevant documents on the web. Compare to the standard web search engines, focused crawler yield good recall as well as good precision by restricting themselves to a limited domain [2].

Focused crawlers look for a subject, usually a set of keywords dictated by search engine, as they traverse web pages. Instead of extracting so many documents from the web without any priority, a focused crawler follows the most appropriate links, leading to retrieval of more relevant pages and greater saves in resources. They usually use a best-first search method called the crawling strategy to determine which hyperlink to follow next. Better crawling strategies result in higher precision of retrieval.

2. GENETIC ALGORITHM

Genetic Algorithms [6] 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. In a typical Genetic Algorithm problem, the search space is usually represented as a number of individuals called chromosomes which make the initial population and the goal is to obtain a set of qualified chromosomes after some generations. The quality of a chromosome is measured by a fitness function (Jaccard in our experiment). Each generation produces new children by applying genetic crossover and mutation operators. Usually, the process ends while two consecutive generations do not produce a significant fitness improvement or terminates after producing a certain number of new generations.

3. THE VECTOR SPACE DOCUMENT MODEL[7]

In the vector space model, we are given a collection of documents and a query. The goal of the query is to retrieve relevant documents out of the collection. In this model, both the documents and the query are represented as vectors.

Once we have document and query vectors, the process of retrieving relevant documents is to compare two vectors together. Given the following terminology, we can compare vectors in several different ways:

Given vectors :

$X=(x_1,x_2,\dots, x_t)$ and

$Y=(y_1, y_2,\dots,y_t)$

Where:

xi - weight of term i in the document and

yi - weight of term i in the query

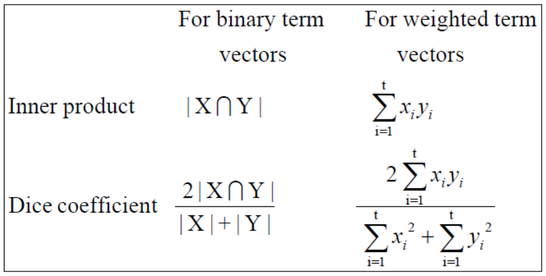
Additionally, for binary weights, let:

|X| = number of 1s in the document and

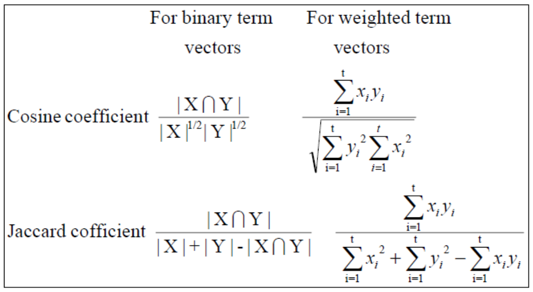
|Y| = number of 1s in query

The following are some methods for comparing vectors:

Figure 1: various similarity coefficients



	For binary term vectors	For weighted term vectors
Inner product	$\lvert X \cap Y \rvert$	$\sum_{i=1}^{t} x_i y_i$
Dice coefficient	$\frac{2\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$	$\frac{2\sum_{i=1}^{t} x_i y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2}$



	For binary term vectors	For weighted term vectors
Cosine coefficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert^{1/2} \lvert Y \rvert^{1/2}}$	$\frac{\sum_{i=1}^{t} x_i y_i}{\sqrt{\sum_{i=1}^{t} y_i^2 \sum_{i=1}^{t} x_i^2}}$
Jaccard cofficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert - \lvert X \cap Y \rvert}$	$\frac{\sum_{i=1}^{t} x_i y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2 - \sum_{i=1}^{t} x_i y_i}$

From the above we have used jaccard similarity coefficient for comparing document and query vector.

4. EXPERIMENT WORK AND EMPIRICAL RESULTS

In our experiment we have selected few queries initially and retrieved first 1o documents from the Google search engine. This we have done for generating chromosomes and extracts the keyword with the highest frequency from each of these pages. These keywords are arranged in the same order as their associated documents were downloaded in an array with n elements which is chromosome length. The length of chromosome is a matter of choice and depends upon number of keywords collectively from the 10 documents. We have chosen chromosome length to be of 21.

Average relevancy of each set of document for a single query was calculated using jaccard quotient as fitness function and applying the selection, crossover and mutation operation. We have selected roulette function for selection of fittest chromosomes after each generation.

$$fitness(d_j) = \sum_{k=1}^{n} \left[\frac{|d_j \cap d_q|}{|d_j \cup d_q|} \right] \qquad (1)$$

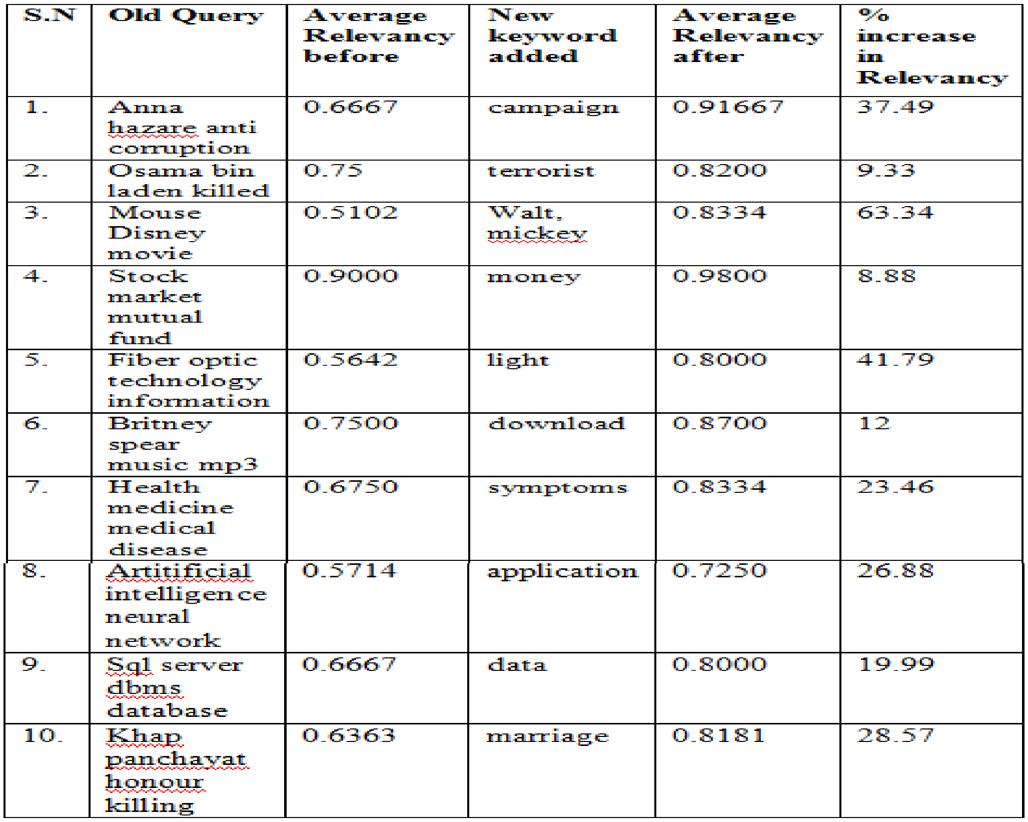
Jaccard Coefficient

Here d_j is the any document and d_q is query document. Both are represented as vector of n terms. For each term appearing in the query if appears in any of the 10 documents in the set a 1 was put at that position else 0 was put.

After finding average relevancy for a set of documents for a single query the same query was extended using one or two new keywords that was mostly found in almost all of the downloaded pages. The results were quite promising and are presented below:

1. Probability of crossover Pc=0.8
2. Probability of mutation Pm=0.01

Table 1: Average relevancy using jaccard coefficient



S.N	Old Query	Average Relevancy before	New keyword added	Average Relevancy after	% increase in Relevancy
1.	Anna hazare anti corruption	0.6667	campaign	0.91667	37.49
2.	Osama bin laden killed	0.75	terrorist	0.8200	9.33
3.	Mouse Disney movie	0.5102	Walt, mickey	0.8334	63.34
4.	Stock market mutual fund	0.9000	money	0.9800	8.88
5.	Fiber optic technology information	0.5642	light	0.8000	41.79
6.	Britney spear music mp3	0.7500	download	0.8700	12
7.	Health medicine medical disease	0.6750	symptoms	0.8334	23.46
8.	Artitificial intelligence neural network	0.5714	application	0.7250	26.88
9.	Sql server dbms database	0.6667	data	0.8000	19.99
10.	Khap panchayat honour killing	0.6363	marriage	0.8181	28.57

5. CONCLUSION & FUTURE WORK

We have conducted several experiments using number of initial keyword as shown above in the table. But we selected only the first 10 pages out of the google search result. This can be extended for 30-50 pages for a precise calculation of efficiency. Further we have shown only the result for Pc=0.8 and Pm=0.01. For various values of Pc and Pm results can be obtained. In conclusion, although the initial results are encouraging, there is still a long way to achieve the greatest possible crawling efficiency.

The work can be extended by using other similarity coefficients like dice and cosine and comparing the result.

6. REFERENCES

[1] D. Michelangelo, C. Frans, L.Steve, C. Lee , G.Marco(2000) Focused Crawling using Context Graphs: Proceedings of the 26th International Conference on Very Large Databases, pp. 527–534.

[2] E. Martin Ester, G.Matthias, K. Hans-Peter Kriegel, Focused Web Crawling(2001): A Generic Framework for Specifying the User Interest and for Adaptive Crawling Strategies : Proceedings of the 27th International Conference on Very Large Database, pp.633-637.

[3] F. Menczer, G. Pant, P. Srinivasan and M. Ruiz(2001) Evaluating Topic-Driven Web Crawlers: In Proceedings of the 24th annual International ACM/SIGIR Conference, pp.531-535.

[4] J. Holland(1975),Adaption in natural and artificial systems : University of Michigan Press,

[5] D. E. Goldberg(1989),Genetic Algorithms in Search, Optimization, and Machine Learning: Addison-Wesley

[6] Shokouhi, M.; Chubak, P.; Raeesy, Z((2005), Enhancing focused crawling with genetic algorithms, Vol: 4-6, pp.503-508

[7] Information retrieval.pdf