Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm

Vikas Thada
Research Scholar
Department of Computer Science and Engineering
Dr. K.N.M University,Newai, Rajasthan, India

Dr Vivek Jaglan
Department of Computer Science and Engineering
Amity University ,Gurgaony,Haryanae, India

Abstract- A similarity coefficient represents the similarity between two documents, two queries, or one document and one query. The retrieved documents can also be ranked in the order of presumed importance. A similarity coefficient is a function which computes the degree of similarity between a pair of text objects. There are a large number of similarity coefficients proposed in the literature, because the best similarity measure doesn't exist (yet !). In this paper we do a comparative analysis for finding out the most relevant document for the given set of keyword by using three similarity coefficients viz Jaccard, Dice and Cosine coefficients. This we perform using genetic algorithm approach. Due to the randomized nature of genetic algorithm the best fitness value is the average of 10 runs of the same code for a fixed number of iterations.The similarity coefficient for a set of documents retrieved for a given query from Google are find out then average relevancy in terms of fitness values using similarity coefficients is calculated. In this paper we have averaged 10 different generations for each query by running the program 10 times for the fixed value of Probability of Crossover Pc=0.7 and Probability of Mutation Pm=0.01. The same experiment was conducted for 10 queries.

Keywords – algorithm, coefficient, genetic, jaccard, relevancy,dice,cosine

I. 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. The ideal focused crawler retrieves the maximal set of relevant pages while simultaneously traversing the minimal number of irrelevant documents on the web. [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.

II. 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.

The process of Genetic Algorithm is as follows:

a. Initialize Population
b. Loop

    i. Evaluation
    ii. Selection
    iii. Reproduction
    iv. Croosover
    v. Mutation
c. Convergence

The initial population is usually represented as a number of individuals called chromosomes. 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.

III. EXPERIMENT WORK AND EMPIRICAL RESULTS

In our experiment we have selected few queries initially and retrieved first 10 documents from the Google search engine. This we have done for generating chromosomes and extract 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,Dice and Cosine similarity coefficients as fitness function and applying the selection, crossover and mutation operation. The complete coding has been done in Matlab software R2009b version. We have selected roulette function or selection of fittest chromosomes after each generation. The three coefficients are shown in table 1.

Table 1. The three similarity coefficients

Similarity Coefficient (X,Y)	**Actual Formula**
Dice Coefficient	$2\frac{\lvert X\cap Y\rvert}{\lvert X\rvert+\lvert Y\rvert}$
Cosine Coefficient	$\frac{\lvert X\cap Y\rvert}{\lvert X\rvert^{1/2}.\lvert Y\rvert^{1/2}}$
Jaccard Coefficient	$\frac{\lvert X\cap Y\rvert}{\lvert X\rvert+\lvert Y\rvert-\lvert X\cap Y\rvert}$

In the table X represents any of the 10 documents and Y represents the corresponding query. 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. The fitness function returns values in the range [0,1].

Table 2: Best fitness values for different queries

Query	**Jaccard**	**Dice**	**Cosine**
Anna hazare anti corruption	0.245556	0.392462	0.498964

Osama bin laden killed	0.254118	0.414344	0.500266
Mouse disney movie	0.190834	0.308178	0.43258
Stock market mutual fund	0.257926	0.417308	0.478768
Fiber optic technology information	0.288336	0.453104	0.549814
Britney spear music mp3	0.267144	0.372908	0.510478
Health medicine medical disease	0.263516	0.381148	0.517844
Artitificial intelligence neural network	0.240868	0.392446	0.513668
Sql server dbms database	0.23856	0.397308	0.464326
Khap panchayat honour killing	0.247892	0.387372	0.492134

The parameters of GA used for finding the results in table 2 are

1. Probability of crossover Pc=0.7
2. Probability of mutation Pm=0.01
3. Number of iterations=150

Graphical comparison is shown in figure 1.

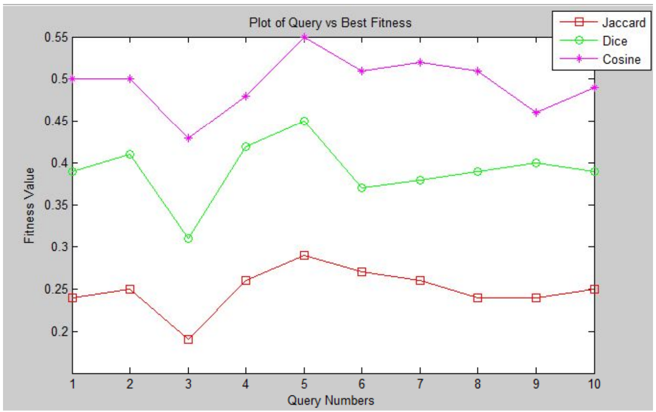

Fig 1 Comparison of similarity coefficients for fitness value

IV.CONCLUSION AND FUTURE WORK

We have conducted several experiments using number of initial keyword as shown above in the table 1. From the table 2 and figure 1 it is clearly visible that best fitness values were obtained using the Cosine similarity coefficients followed by Dice and Jaccard. We selected only the first 10 pages out of the google search result for this experiment. This is being extended in the future research for 30-35 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 will be obtained. In conclusion, although the initial results are encouraging, there is still a long way to achieve the greatest possible crawling efficiency.

REFERENCES

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

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

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

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

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

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

[7] Information retrieval.pdf, Google