International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Use of Genetic Algorithm to Find Average Number of Generations to Converge for Web Retrieved Documents Using Jaccard Similarity Coefficient

[1]Mr. Vikas Thada, [2]Dr. Vivek Jaglan
[1]Research Scholar(CSE), . [2]*Asst.Prof(CSE)*
[1]Dr.K.N.Modi University, [2]Amity University
[1] Newai, [2]Gurgoan, India

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. Due to the randomized nature of genetic algorithm we show that generation number for convergence value of 1 is not always the same. The similarity coefficient for a set of documents retrieved for a given query from Google are find out then average relevancy 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=0.10. The same experiment is conducted for 10 queries.*

Keywords: *Focused, crawler, retrieval, relevancy, similarity, algorithm, coefficient, genetic, jaccard, relevancy.*

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 followsthe 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.

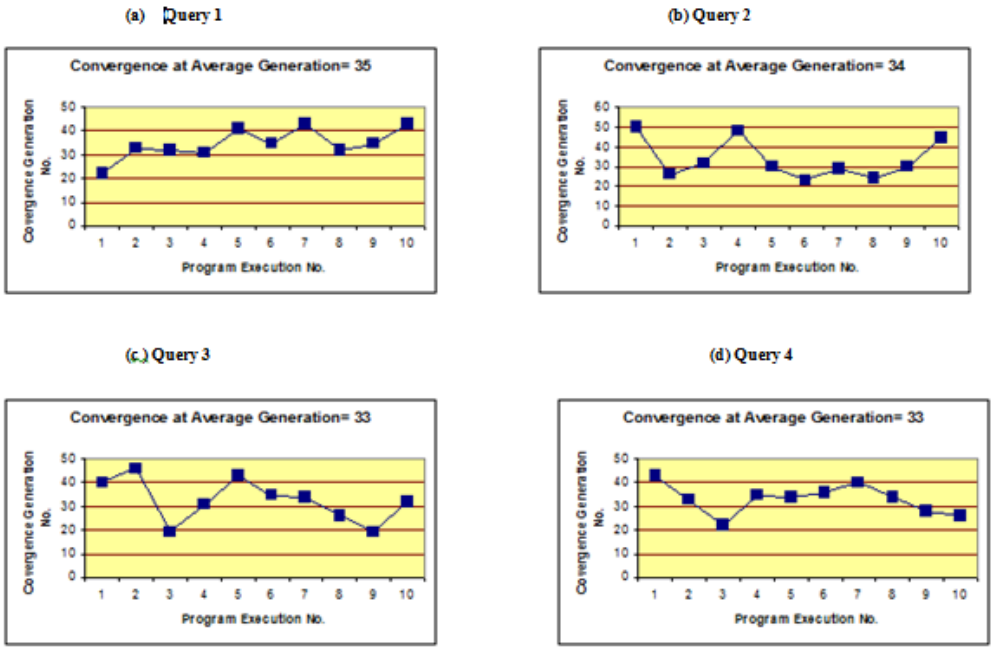
Average relevancy of each set of document for a single query was calculated using Jaccard quotient as fitness function and applying the selection, crossover and mutation operation. We have selected roulette function or selection of fittest chromosomes after each generation.

$$\text{Fitness}(d_j)=\sum_{k=1}^{k=n}\left[\frac{d_j \cup d_q}{d_j \cap d_q}\right] \quad \textbf{(1)}$$

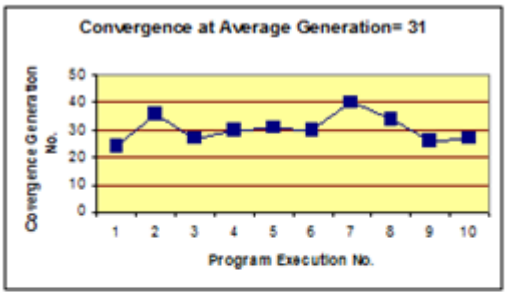
The fitness function is shown above as equation 1. Here dj is the any document and dq is query document. Both are represented as vector of n terms. For each term appearing in the query if appears in any of the 10 documents in the set a 1 was put at that position else 0 was put. The fitness function returns values in the range [0,1].

1. Probability of crossover Pc=0.7
2. Probability of mutation Pm=0.10

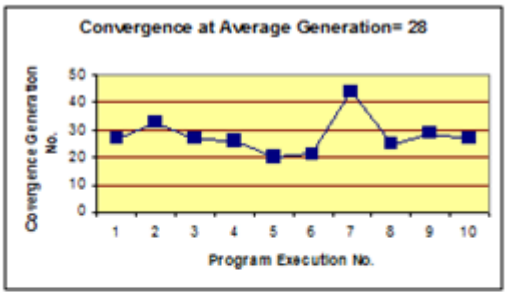
Figure 1: Average Generation For Convergence Value= 1 For Different Queries



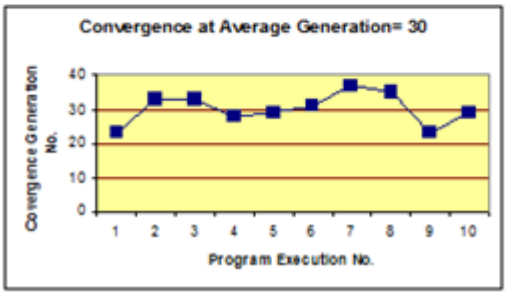
(e) Query 5



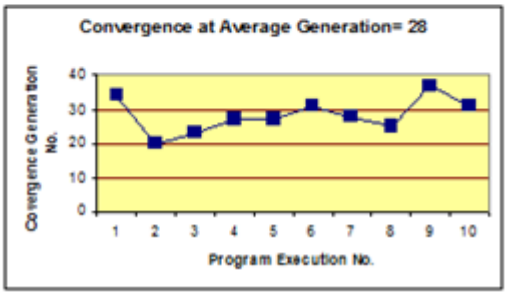
(f) Query 6



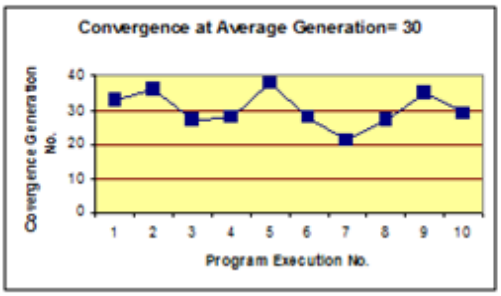
(g) Query 7



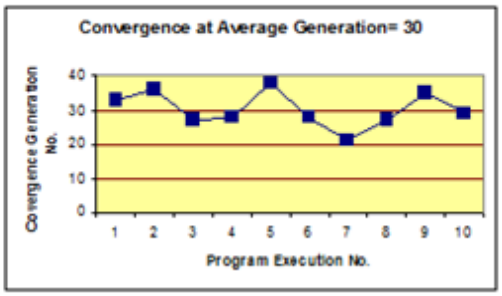
(h) Query 8



(i) Query 9



(j) Query 10



The graphs shown above were plotted using the data shown below in the table. For every query results were obtained by feeding the input query and documents through algorithm implemented in C++.

S.N	Query	10 different generations for convergence value=1										Average
1.	Anna hazare anti-corruption	22	33	32	31	41	35	43	32	35	43	35
2.	Osama bin laden killed	27	26	32	48	30	23	29	24	30	45	34
3.	Mouse Disney movie	40	46	19	31	43	35	34	26	19	32	33
4.	Stock market mutual fund	43	33	22	35	34	36	40	34	28	26	33
5.	Fiber optic technology information	24	36	27	30	31	30	40	34	26	27	31
6.	Britney spear music mp3	27	33	27	26	20	21	44	25	29	27	28
7.	Health medicine medical disease	23	33	33	28	29	31	37	35	23	29	30
8.	Artitificial intelligence neural network	34	20	23	27	27	31	28	25	37	31	28
9.	Sql server dbms database	33	36	27	28	38	28	21	27	35	29	30
10.	Khap panchayat honour killing	30	35	36	28	29	35	37	29	21	23	30

Table 1
Average Generation For Convergence Value= 1

IV. Conclusion & Future Work

Several experiments were carried out with different set of query words as shown above in the table 1. The algorithm was encoded in C++ programming language. The database size for query returned pages were restricted to only 10 pages. This can be extended for 30-50 pages for a precise calculation of efficiency. For varying probability of Pc and Pm different results can be obtained. Although the initial results are encouraging, there is still a long way to achieve the greatest possible crawling efficiency.

As a future extension of this research work other similarity coefficients like dice and cosine can be used and result can be compared. Further automated process of extraction of top keywords can be done.

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 27th 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

Acknowledgement

I thank Google for fast and efficient search and many other research papers that we've studied to help do this research work. Indirectly I thank to all my colleages at my worksplace, my supervisor and above all my family for constant support in all my endeavours.