International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 3.45 (SJIF-2015), e-ISSN: 2455-2584
Volume 3, Issue 3, March-2017

# EFECTIVENESS OF FOCUS CRAWLING USING GENETIC ALGORITHM

[1]Ravindra R. Rathod, [2]Suchindar Shingh

[1]*M.Tech Student, Electronics & Communication, SSIET, Derabassi Punjab, Punjab,*

[2]*Electronics & Communication (Associate Professor), SSIET, Derabassi, Punjab*

***Abstract — GA are robust, efficient and optimized methods that are used in getting solutions to number of search and optimization problems. In this paper, we use a genetic algorithm with focused crawling for improving the accuracy of similarity measure. As information has been increasing rapidly, it is very difficult to get information as per the user satisfaction. A focused crawler is a web crawler which gives relevant pages according to users query.Genetic Algorithms (GA) are optimization algorithms inspired by the Darwin's theory of natural evolution and survival of fittest.***

***Keywords- web crawler, relevancy Crawling, focused crawling, Genetic Algorithm,***

## I. INTRODUCTION

The search results of the entered keyword sometimes might not display the specific web pages which might be due to the lack of the search method. Thus, there were several researches tackling these problems that had an effect on the accuracy of the system in order to find the best solutions to increase the efficiency and accuracy of the system. There is enormous amount of information available on the internet in the form of text, image, audio and video. As the major content available from the web world is in the form of text so to retrieve the relevant text is still a challenging job for any focus crawler .The user mostly types his query as text in the search engine. Focused crawlers looks for a subject, usually a set of keywords dictated by search engine, as they traverse the 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. In this study, results showed some improvements in focus crawler performance using genetic algorithm, through implementing some queries in order to obtain relevant information, sorting such queries depending on similarity function.

## II. RELATED WORKS TO FOCUSED CRAWLERS

Bangorn Klabbankoh and Ouen Pinngern [4] analyzed vector space model to boost information retrieval efficiency. In vector space model, IR is based on the similarity measurement between query and documents. Documents with high similarity to query are judge more relevant to the query and will be retrieving first. Testing result will show that information retrieval with 0.8 crossover probability and 0.01 mutation probability provide the maximum precision while 0.8 crossover probability and 0.3 mutation probability provide the maximum recall. The information retrieval efficiency measures from recall and precision.There are several researches that used genetic algorithm with focus crawler to optimize the user query. Chakrabarti et al.[1] introduced focused crawler for the first time. The crawler described in their article [2], the user picks a subject from a pool of hierarchically structured example documents. The program learns by studying the examples, and generates subject models. These models are used to classify web pages. The link structure is also considered by the crawler to discover hubs. Hubs are described by Kleinberg as high-quality lists that guide users to recommended authorities, and authorities are prominent sources of primary content on a topic. Links from hubs can be relevant even though the text on the hub page itself does not appear to be relevant. Semiautomatic web resource discovery using ontology-focused crawling [20] To use which parameters depends on the appropriateness that what would user like to retrieve for. In the case of high precision documents prefer, the parameters may be high crossover probability and low mutation probability. While in the case of additional relevant documents prefer the parameters may be high mutation probability and lower crossover probability in information retrieval. A tested database consisted documents taken from student's projects. Beginning experiment indicated that precision and recall are invert.

## III. GENETIC ALGORITHM

We will be talking about deception, population sizing studies, the role of parameters and operators, building block mixing, and linkage learning. These studies are motivated by the desire of building better GAs, algorithms that can solve difficult problems quickly, accurately, and reliably. It is therefore a theory that is guided by practical matters.

A genetic algorithm is a search procedure inspired by principles from natural selection and genetics. It is often used as an optimization method to solve problems where little is known about the objective function. The operation of the genetic algorithm is quite simple. It starts with a population of random individuals, each corresponding to a particular candidate

 

solution to the problem to be solved. Then, the best individuals survive, mate, and create offspring, originating a new population of individuals. This process is repeated a number of times, and typically leads to better and better individuals of current genetic algorithm theory. This theory is centered arround the notion of a building block.

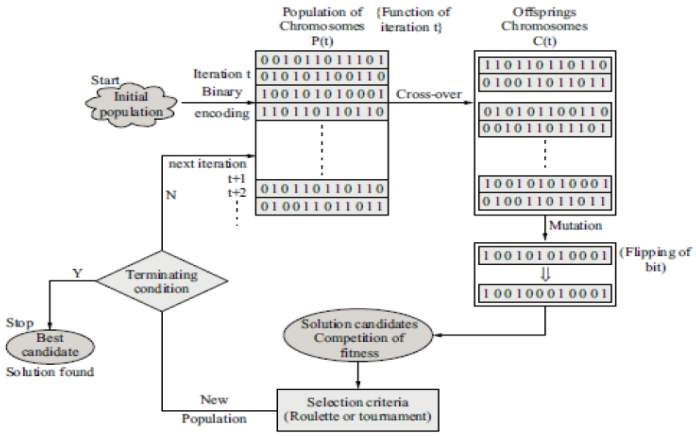


***Figure 1. Flow of genetic algorithms***

### 3.1 Genetic Algorithms in Information Retrieval

Genetic algorithms are adaptive heuristic search algorithm . Genetic algorithm is based on evolutionary ideas of natural selection and genetics. Genetic algorithm is often used to solve problems and looking for best solution. In genetic algorithm query entered by user and text documents of news articles stored in data repository represented as chromosome. Each chromosome has specific matching score. By deciding some threshold value we compare the threshold value with matching score and if matching score is greater than threshold value then that chromosomes will be fittest. And if matching score is less than threshold value then that chromosomes will be taken to generate the next generation. The next generation is generated by using genetic operators such as crossover, mutation Genetic algorithms are adaptive heuristic search algorithm. Genetic algorithm is based on evolutionary ideas of natural selection and genetics. Genetic algorithm is often used to solve problems and looking for best solution. In genetic algorithm query entered by user and text documents of news articles stored in data repository represented as chromosome. Each chromosome has specific matching score. By deciding some threshold value we compare the threshold value with matching score and if matching score is greater than threshold value then that chromosomes will be fittest. And if matching score is less than threshold value then that chromosomes will be taken to generate the next generation. The next generation is generated by using genetic operators such as crossover, mutation and reinsertion.

### 3.2 Fitness Measure

The score of a fitness function is a numerical value that indicates how well the particular solution solves the problem. The score is the fitness of the individual solution. It represents how well the individual adapts to the environment. The task of the GA is to discover solutions that have high fitness values among the set of all possible solutions. We use Tanimoto coefficient as fitness function for our genetic algorithm.

### 3.3 Chromosome Representation

Both documents and queries are represented by vector. A document vector with n keywords and a query vector with m query terms can be represented as
Doc = (term1, term2 , term3,…. termn)
Query = (qterm1, qterm2, qterm3 , …qtermm)
For example, user enters a query into Google search engine that could retrieve 10 documents. Then user gets frequently used words form a tool called Text analyser. The length of chromosome depends on number of keywords of documents retrieved from user query.

### 3.4 Selection

Selection is the process in which chromosomes is selected for next step or generation in genetic algorithm based on fitness value of chromosomes. Poor chromosome or lowest fitness chromosome selected few or not at all.

### 3.5 Crossover

Crossover is the genetic operators that mix two chromosomes together to form new offspring. Crossover occurs only with

some probability Pc (crossover probability). GA's construct a better solution by mixture good characteristic of chromosomes together. Higher fitness chromosomes have an opportunity to be selected more than the lower ones, so good solution always alive to the next generation.
For example, two chromosomes are crossover between position 9 and 14.

Doc1 ={0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 1 0 0 0}
Doc2 ={0 0 0 0 0 1 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0}
The resulting crossover yields two new chromosomes.
Doc1 ={0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0}
Doc2 ={0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0}

**3.6 Mutation**

Mutation involves the modification of the values of each gene of a solution with some probability Pm (mutation probability). In accordance with changing some bit values of chromosomes, give the different breeds. For example randomly at position 10 apply mutation.
Doc1 ={0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 1 0 0 0}
Result {0 0 0 0 1 0 0 0 0 0 1 0 1 1 0 0 0 1 0 0 0}

## IV. EMPIRICAL RESULTS

Table 1. shows initial keywords and the terms added with the help of genetic algorithm. We entered the old and new keywords in addition to old keywords into Google, and calculated the average relevance based on the10 documents returned. As shown in our experiment , after adding new set of keywords documents achieved a higher relevance score. For the first sample, the task is to search for news about vyapam scam. After downloading about 10 pages, Genetic algorithm added "madhyapradesh" to its initial set. This word was chosen because in most of the downloaded pages there was about vyapam scam in Madhya Pradesh. This experimentation tests for 10 queries with fitness functions Tanimoto.

## V. CONCLUSIONS

The paper discusses fundamentals of Web Information Retrieval. The paper has also deal with how GA can be used in the field of Web-IR and can efficiently help in finding relevant documents. Paper has also discussed plenty of research work done earlier in the field of Web-IR with GA. An experimental setup using Tanimoto coefficient for Web-IR and finding relevant documents has also been discussed. Although the initial results are encouraging, there is still a long way to achieve the greatest possible crawling efficiency.

***Table 1. A.R=Average Relevance by Tanimoto fitness function***

| Old Keyword | New Added Term | Average Relevence With Old Keyword | Average relevence With add new Keyword | Percentage Improvement |
|---|---|---|---|---|
| Vyapam, scam | **Madhya Pradesh** | 0.6255 | 0.7856 | 26 |
| Aam admi party | **Arvind, kejriwal** | 0.8135 | 0.9243 | 14 |
| France, French, Revolution, History | **Napoleon** | 0.5432 | 0.6489 | 19 |
| Micheal Jaekson music mp3 | **Download** | 0.6842 | 0.8329 | 22 |
| Mouse Disney movie | **Walt, mickey** | 0.3515 | 0.5400 | 54 |

**REFERENCES**

[1] MPS Bhatia, Akshi Kumar Khalid, "A Primer on the Web Information Retrieval Paradigm" Journal of Theoretical and Applied Information Technology, p. 657-662.

[2] Gautam Pant, Padmini Srinivasan1, and Filippo Menczer, "Crawling the Web" in procd Web Dynamics pp.153-178, 2004.

[3] Anshika Pal, Deepak Singh Tomar, S.C. Shrivastava, "Effective Focused Crawling Based On Content And Link Structure Analysis" International Journal of Computer Science and Information Security, Vol. 2, No. 1, June 2009.

[4] Michelangelo Diligenti, Frans Coetzee, Steve Lawrence, C. Lee Giles, Marco Gori, " Focused Crawling using Context Graphs," Proceedings of the 26th VLDB Conference, Cairo, p. 527–534, 2000.

[5] S. Chakrabarti, M. van der Berg, and B. Dom, "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery," in Proc. 8th International World- Wide Web Conference, p. 545-562, 1999.

[6] J. Cho, H. Garcia-Molina, and L. Page, "Efficient Crawling Through URL Ordering," In Proceedings of the Seventh International World Wide Web Conference. Volume 30, April, P. 161-172, 1998.

[7] B. Klabbankoh and O. Pinngern, "Applied genetic algorithms in information retrieval," IJCIM, vol. 7, no. 3, December 1999.

[8] Z. Gao, Y. Du, L. Yi, Q. Peng, Y. Yang ," Incrementally Updating Concept Context Graph (CCG) for Focused Web Crawling Based on FCA" In proc. Asia-Pacific Conference on Information Processing, vol. 2, p.40-43, 2009.

[9] Ahmed Ghozia , Hoda Sorour and Ashraf Aboshosha," Improved Focused Crawling Using Bayesian Object Based Approach," In proceeding a Radio Science Conference, p.1 – 8, 2008.

[10] Qu Cheng, Wang Beizhan, Wei Pianpian, "Efficient Focused Crawling Strategy Using Combination of Link Structure and Content Similarity" Proceedings of IEEE International Symposium on IT in Medicine and Education. vol.2, July, p.797 – 802, 2003.

[11] Bing Liu, Chee Wee Chin, Hwee Tou Ng. "Mining Topic-Specific Concepts and Definitions on the Web" in proceeding WWW, May 20-24, Hungary, 2003 .

[12] T. Peng, W.L. Zuo and Y.L. Liu "Genetic Algorithm For Evaluation Metrics In Topical Web Crawling" Computational Methods Springer in the Netherlands, pp- 1203–1208, 2006.

[13] J. J. Gregory Caporaso William A. Baumgartner, Jr. Hyunmin Kim, Zhiyong Lu Helen L. Johnson Olga Medvedeva Anna Lindemann, Lynne M. Fox Elizabeth K. White K. Bretonnel Cohen Lawrence Hunter, "Concept Recognition, Information Retrieval, and Machine Learning in Genomics Question-Answering" in proc. TREC Proceedings (723), November, 2006.

[14] Soumen Chakrabarti, Kunal Punera, Mallela Subramanyam "Accelerated Focused Crawling through Online Relevance Feedback" WWW2002, May 7-11, Honolulu, Hawaii, USA 2002.

[15] Blaž Novak "A Survey Of Focused Web Crawling Algorithms" Publication Year, multiconference is 2004, 12-15 Oct 2004, Ljubljana, Slovenia.

[16] Yuxin Chen, Edward A. Fox et. al "A Novel Hybrid Focused Crawling Algorithm to Build Domain-Specific Collections" Virginia Polytechnic Institute & State University Blacksburg, VA, USA pp- 85, 2007

[17] Ahmed A. A. Radwan, Bahgat A. Abdel Latef, Abdel Mgeid A. Ali, and Osman A. Sadek "Using Genetic Algorithm to Improve Information Retrieval Systems" World Academy of Science, Engineering and Technology 17 2006 ISSN 2070-3724.

[18] Jialun Qin & Hsinchun Chen "Using Genetic Algorithm in Building Domain-Specific Collections An Experiment in the Nanotechnology Domain" Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05), Volume 04 IEEE Computer Society Washington, 2005.

[19] Hsinchun Chen, Yi-Ming Chung, and Marshall Ramsey "A Smart Itsy Bitsy Spider for the Web" Journal of the American Society for Information Science pp 604–618, 1998.

[20] Alessandro Micarelli, Fabio Gasparetti "Adaptive focused crawling" Lecture Notes in Computer Science the adaptive web methods and strategies of web personalization section Adaptation technologies pp 231-262, 2007.

[21] Chain Singh, Ashish Kr. Luhach, Amitesh Kumar " Improving Focused Crawling with Genetic Algorithms" International Journal of Computer Applications, *2013*

[22] Milad shokouhi, Pirooz Chubak, Zaynab Raeesy," Enhancing Focused Crawling with Genetic Algorithms," Information Technology: Coding and Computing, Volume 2, Issue, 4-6 April P. 503 – 508, 2005.

[23] Knut magne risvik and Rolf michelsen, "Search Engines and Web Dynamics," in proceeding of computer networks volume 39, Issue 3, 21 June, P. 289-302, 2002.

[24] Chakrabart S., van den Berg, M. Dom, "Distributed Hypertext Resource Discovery through Examples" In Proceedings of the 25th International Conference on Very Large Data Bases. P. 375 – 386, 99.