Improving Focused Crawling with Genetic Algorithms

Chain Singh
DCE,Gurgaon
Farrukhnagar,
Gurgaon

Ashish Kr. Luhach
DCE,Gurgaon
Farrukhnagar,
Gurgaon

Amitesh Kumar
DCE,Gurgaon
Farrukhnagar,
Gurgaon

ABSTRACT
The Web, containing a large amount of useful information and resources, is expanding rapidly. Web crawlers are one of the most crucial components in search engines and their optimization would have a great effect on improving the searching efficiency. Focused Crawlers can selectively retrieve Web documents relevant to a specific domain to build collections for domain-specific search engines. In this paper, we use a genetic algorithm with focused crawling for improving its crawling performance. Expands initial keywords by using a genetic algorithm for focused crawling. The results showed that our approach could build domain-specific collections with higher quality than traditional focused crawling techniques.

General Terms
Focused crawling, keyword enhance

Keywords
Crawling, focused crawling, Genetic Algorithm, web crawler

1. INTRODUCTION
A Web crawler is a key component inside a search engine [1]. It can traverse the Web space by following Web page's hyperlinks and storing the downloaded Web documents in local repositories that will later be indexed and used to respond to the user's queries efficiently [2]. However, with the huge size and explosive growth of the Web, it becomes more and more difficult for search engines to provide effective services to end-users. Moreover, such a large collection often returns thousands of result documents in response to a single query. It is impossible for major search engines to update their collections to meet such rapid growth. As a result, end-users often find the information provided by major search engines not comprehensive or out-of date.

To address the above problems, focused crawler were introduced. A focused crawler or topical crawler is a web crawler that attempts to download only web pages that are relevant to a pre-defined topic or set of topics. Focused crawling was first introduced by Chakrabarti et al.[7]. Most focused crawlers use the content of traversed pages to determine the next hyperlink to crawl. They use a similarity function to find the most similar page to the initial keywords that is already downloaded and crawl the most similar one in the next step. These similarity functions use information retrieval techniques [20] to assign a weight to each page so that the page with the highest weight is more likely to have the most similar content.

2. Related Works to Focused Crawlers
Chakrabarti et al. seem to introduce focused crawling for the first time. In the crawler described in their article [2], the user picks a subject from a pool of hierarchically structured example documents. The program learns the subjects 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. Semi automatic web resource discovery using ontology-focused crawling [20]

The crawler described by Chakrabarti et al. [2] uses example documents and machine learning principles. One difference with Diligenti's [1] crawler is that it generates a context graph that describes the link structure around all the seed documents. Diligenti's crawler only focuses on web pages. Ester's crawler and other crawlers that use a static initial set of example documents for classification are very dependent on the quality of the initial training data. Sizov et al. have built a focused crawler that aims to overcome the limitation of the initial training data there are some other experiments which measure the similarity of page contents with a specific subject using special metrics and reorder the downloaded URLs for the next crawl [3] or even evaluate a learning scheme for identifying which URL the spider should crawl next in order to increase the efficiency in topic specific web resource discovery [25]. Bing Liu et al combined the crawling strategy with clustering concepts [14]. For each topic they first retrieve a specific number of top weighted retrieved pages from Google for that topic and then extract some other keywords from them.

3. Genetic Algorithm
Genetic algorithms (GA) are search algorithms based on the principle of natural selection and genetics. GA operates on a population of potential solutions applying the principle of the survival of the fittest to produce better and better approximation to the solution of the problem that GA is trying to solve. At each generation, a new set of approximations is created by the process of selecting individuals according to their level of fitness value in the problem domain and breeding them together using the operators borrowed from the genetic process performed in the nature, i.e. crossover and mutation. This process leads to the evolution of populations of individuals that are better adapted to their environment than the individuals that they were created from, just as it happens in natural adaptation.

3.1 Genetic Algorithms in Information Retrieval
Genetic algorithms (GA) are search algorithms based on the principle of natural selection and genetics. GA operates on a population of potential solutions applying the principle of the survival of the fittest to produce better and better approximation to the solution of the problem that GA is trying to solve. Until 1980 there were no serious attempts in the field of applying genetic algorithms in information retrieval. Raghavan and Aggarwal were the first ones who introduced

the subject by trying to optimize document clustering by the means of genetic algorithms [26]. One year later, Gordon proposed a new approach for improving the document descriptors. In his experiments documents were represented by an array of keywords which evolved over time by natural selection and genetic operators, and the final results were generally proved to be the best string of keywords describing the document [27]. The trend was followed by Yang et al. for improving the weights of keywords associated with the document topic [28]. Petry et al. applied genetic algorithms to increase the functionality of retrieving data from a weighted indexed collection of documents, by modifying weights of query terms.

3.2 Fitness Evaluation

Fitness function is a performance measure or reward function which evaluate how good each solution is. Result from fitness functions are interval 0 to 1. By 1.0 means document and query is sameness. Values near 1.0 mean documents and query are more relevant and values near 0.0 mean documents and query are less relevant. We use Jaccard coefficient as fitness function for our genetic algorithm. We use binary term vector of Jaccard coefficient.

3.3 Chromosome Representation

Both documents and queries are represented by vector. A document vector (Doc) 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)

We use binary term vector. For example, user enters a query into Google search engine that could retrieve 10 documents. These documents are

Query = {Khap, Panchayat, Honour, Killing}

Doc1 = {Honour, Khap, Killing, Panchayat}

Doc2 = {Killing, Decide, Goverment}

.

.

.

Doc10={Killing, Honour, Government, Session, Monsoon}

All keywords of these documents can be arranged in the ascending order as

Against, Boys, Caste, Couples, Court, Decide, Family, Girls, Government, Haryana, Honour, Intra, Khap, Killing, Law, Marriage Monsoon, Panchayat, Session, Union, and Village.

Encode in the chromosome representation as

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}

.

.

Doc10={00 0 0 0 0 0 0 1 0 1 0 0 1 0 0 1 0 1 0 0}

These chromosomes are called initial population that feed into genetic operator process. The length of chromosome depends on number of keywords of documents retrieved from user query. From our example the length of each chromosome is 21 bits.

3.4 Selection

After we evaluate population's fitness, the next step is chromosome selection. Selection embodies the principle of 'survival of the fittest'. Satisfied fitness chromosomes are selected for reproduction. Poor chromosomes or lower fitness chromosomes may be selected a 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 6 and 13.

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 1 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 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 0 0 0}

4. Empirical Results

Table 1 depicts initial keywords and the terms added given by genetic algorithm, written in bold. We entered the old and new keywords (old keywords plus the new term added by genetic algorithms functions) into Google, and calculated the average relevance based on the10 first pages returned for each instance. As it is shown, the new set of keywords achieved a higher relevance score. For the first sample, the task is to search for news about khap panchayat. After downloading about 10 pages, Genetic algorithm added "marriage" to its initial set. This word was chosen because in most of the downloaded pages there was about marriage in Haryana.

This experimentation tests for 10 queries with fitness functions Jaccard coefficient. Average relevance is defined by the following equation

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

Jaccard coefficient

5. Conclusions

As the size of the Web keeps growing, it has become increasingly important to build high-quality domain specific search engines. This research has proposed a new crawling technique to build domain-specific collections for search engines that incorporate a global search algorithm, Genetic Algorithm, into the crawling process.

In our user study, our proposed Genetic Algorithm with focused crawling built collections with significantly higher quality than did a traditional best-first crawler. Our proposed work apply a genetic algorithm to the focused crawling process to find more relevant Web resources in order to overcome the problems faced by traditional focused crawlers. Our goal is to extend the keyword set for the focused crawling. In this way searching made easy, we find more relevant document or web pages. Results shows that average relevance of document increase upto 50%. When focused crawler having key set with more relevancy then retrieved data also more relevant for local collection of a search engine. It improves the crawling performance.

Table 1: A.R=average relevance

Old Keyword	New Added Term	A.R With Old Keyword	A.R With add new Keyword
Khap Panchayat honour killing	**Marriage**	**0.3081**	**0.3911**
Fiber optics technology information	**Light**	**0.4175**	**0.4729**
Genetic algorithm optimization softcomputing	**Fitness function**	**0.2646**	**0.3466**
Micheal Jaekson music mp3	**Download**	**0.3616**	**0.5208**
Mouse Disney movie	**Walt mickey**	**0.3515**	**0.5400**
Economy finance stock market financial	**News**	**0.2232**	**0.3567**
Health medicine Medical disease	**Symptoms**	**0.2304**	**0.4197**
Sql server dbms database	**Data**	**0.3144**	**0.4321**
Artificial intelligence neural network	**Applicatio n**	**0.4136**	**0.5596**

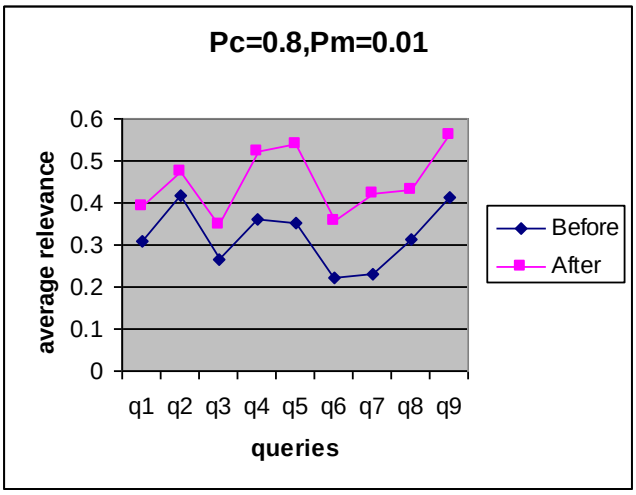

Fig 1

6. Future Work

There is still a lot of work to do for improving the efficiency of the focused crawling. There is also a downside to the algorithm. Some websites repeat one word their abbreviated name frequently. In this case, when the page content is really close to the search subject, the extracted keyword might be irrelevant to the initial keyword set. For example, when we were extracting keyword from a web page for news about the Khap panchayat, we chose the google.com as the initial website and since the term was repeated so many times in almost all the pages that are not relevant to the query but related to that site like news, HT etc. This term was returned by the crawler as the keyword. This problem may be cumbersome to manage but attempts could be made to find solutions. Also, only one of the bits of the fittest chromosome would be added to the initial set and the other bits will be neglected. We intend to implement some simple learning procedures in order to find out which of these bits might be useful in the future. Work of selecting keyword using text analyzer tool is manual that is also a problem. It is also a future work make it automatically and online.

In conclusion, although the initial results are encouraging, there is still a long way to achieve the greatest possible crawling efficiency.

7. References

[1] 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.

[2] 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.

[3] 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.

[4] M. Jamali, H. Sayyadi, B. Bagheri H. and H. Abolhassani, "A Method for Focused Crawling Using Combination of Link Structure and Content Similarity"

In Proceedings of the International Conference on Web Intelligence table of contents p. 753-756, 2006.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

[14] 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 .

[15] 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.

[16] 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.

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

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

[19] 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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] Bangorn klabbankoh, Ouen pinngern ph.d. "applied genetic algorithms in information retrieval" In proc. IEEE, vol-92,pp-702-711, issue-4, nov 2004.

[25] N. Angkawattanawit and A. Rungsawang, "Learnable Crawling: An Efficient Approach to Topic-Specific web Resource Discovery", 2nd international Symposium on communications and Information Technology (ISCIT 2002), October 2002.

[26] V. Raghavan and B. Aggarwal, "Optimal Determination of User-Oriented Clusters: An Application for the Reproductive Plan," in the Proceedings of the Second International Conference on Genetic Algorithms and Their Applications, Cambridge, pp. 241-246, 1987.

[27] M. Gordon, "Probabilistic and Genetic Algorithms for Document Retrieval," Communications of ACM (31:2), 1988, pp. 152-169.

[28] J. Yang, R. Korfhage, and E. Rasmussen, "Query Improvement in Information Retrieval Using Genetic Algorithms: A Report on the Experiments of the TREC Project," in Proceedings of the First Text Retrieval Conference, Washington, National Institute of Standards and Technology (NIST) Special Publication 500-207, March 1993, pp. 31-58,1993.