A Primer on Genetic Algorithm in WebIR

[1]Dr. VikasThada, [2]Mr. UtpalShrivastava

[1,2]Asst.Prof(CSE), Amity University Gurgaon, India
Email: [1]vthada@ggn.amity.edu, [2]ushrivastava@ggn.amity.edu

Abstract: The field of computer science that deals with searching, retrieving and presenting data or information onto the Web and within online databases and also searches the web documents is known as Web Information Retrieval. Web IR can be defined as the application of theories and methodologies from IR to the World Wide Web. It is concerned with addressing the technological challenges facing Information. The utility and ubiquity of web search is making Web Information Retrieval (IR) an increasingly popular research topic. Genetic Algorithms (GA) are optimization algorithms inspired by the Darwin's theory of natural evolution and survival of fittest. GA are robust, efficient and optimized methods which are used in finding solutions to number of search and optimization problems. The paper describes how GA can be utilized in the field of Web Information Retrieval emphasizing specially onto the searching process onto the World Wide Web.

Keywords: WebIR,Rogers-Tanimoto,Genetic, Algorithm, Web, search

I. INTRODUCTION

Tim Berners-Lee and his World Wide Web entered the information retrieval world in 1989 .This event caused a branch that focused specifically on search within this new document collection to break away from traditional information retrieval. This branch is called web information retrieval [22]. Web Information retrieval is the process of searching within a huge World Wide Web document collection for a particular information need (called a query). By all measures, the Web is enormous and growing at a staggering rate, which has made it increasingly intricate and crucial for both people and programs to have quick and accurate access to Web information and services. Thus, it is imperative to provide users with tools for efficient and effective resource and knowledge discovery. Web IR can be defined as the application of theories and methodologies from IR to the World Wide Web. It is concerned with addressing the technological challenges facing Information Retrieval (IR) in the setting of WWW [23]. Notion of relevance is at the center of web information retrieval;That is retrieving documents as par the user need. How much they are related with the input data or query? In fact, the primary goal of an Web IR system is to retrieve all the documents which are relevant to a user query while retrieving as few non-relevant documents as possible. How GA helps in matching the query with the relevant documents and retrieving optimum number of relevant documents is at the core of this paper.

II. OVERVIEW OF GA

Genetic Algorithms [24] 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 (Rogers-Tanimoto 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.

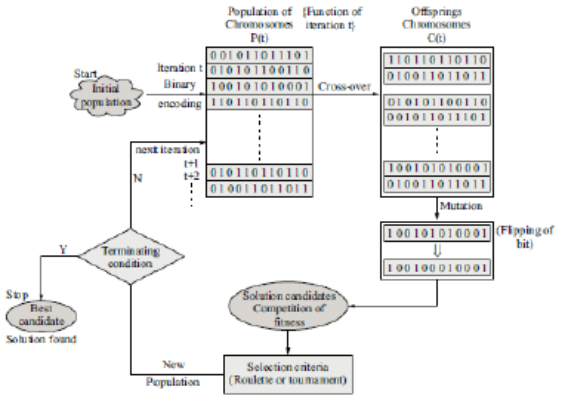

Figure 1 Basic Operation of Genetic Algorithm

III. GA IN WEB IR

In GA implementation, the search space is composed of candidate solutions (called individuals or creatures) to an optimization problem to evolve better solutions. Each

of the candidate is represented by a string of 0 and 1's and is termed as chromosome. Each chromosome has an objective function value, called fitness value. A set of chromosomes together with their associated fitness value is called a population. For associating fitness value to each of the chromosomes some fitness function can be used. Input to this fitness function is the binary representation of the chromosome. This population, at a given iteration of the genetic algorithm, is called a generation. In each generation, the fitness of every individual in the population is evaluated from the current population based on their fitness value and modified by applying selection, crossover and mutation to form a new population. The new population is then used in the next iteration of the algorithm.

GA has been used and is being used in finding the relevant documents from the WWW given some matching examples of relevant and non-relevant documents. Number of proposals for web information retrievals has been proposed by [1][2][3][4][5][6][7][8][9][10][11][12][13][14][15][16][17][18][19][20][21]. Each of them has utilized efficiently GA for the purpose of web search / web information retrieval. Detailed discussion of the above references is out of the scope of this paper but all of them with different fitness function have successfully harnessed the power of GA in web information retrieval using adaptive and non adaptive methods of GA.

IV. THE EXPERIMENTAL SETUP

In general the web documents are encoded in strings of 0's and 1's as shown in the figure 1. The documents have been obtained using some search query. The set of documents also contains some non relevant documents too. In the experimental setup of GA 50 documents have been retrieved for each different query. This set of 50 documents serves as document database. The query is also encoded in terms of string of 0 and 1. Process of experiment is as follows:

1. User enters query into Google.

2. Find keywords from retrieved document with the help of Textalyser or Keyword Density Checker online tool

3.Encode documents retrieved by user query to chromosomes (initial population).

4. Feed the population thus generated into genetic operators viz. selection, crossover, and mutation.

5. Repeat step 4 until maximum generation is reached. At the end we get an optimized query chromosomefor document retrieval.

6. Decode optimize query chromosome to query and retrieve document from database.

7. Check the relevance before and after adding the keyword using Rogers-Tanimoto similarity coefficient and note the improvement in retrieval document

8. Depending upon the maximum relevance the documents are ranked and stored in the document database.

Table 1: Average Relevancy by RogersTanimoto fitness function with Pc = 0.9 and Pm = 0.01

S.N	Old Query	Average Relevancy Before	New Keyword Added	Average Relevancy After	% Increase In Relevancy
1.	**Anna hazareanti corruption**	0.6102	Campaign	0.9334	66.34
2.	**Osama bin laden killed**	0.75	Terrorist	0.8800	12.33
3.	**Mouse disney movie**	0.7667	Walt, mickey	0.8566	43.49
4.	**Delhi gang rape**	0.7650	december	0.89334	21.46
5.	**Search engine optimization**	0.6542	website	0.8000	25.79
6.	**Britney spear music mp3**	0.7514	Download	0.8750	26.88
7.	**Aamaadmi party**	0.9000	Arvind, kejriwal	0.9800	8.88
8.	**Bomb blast boston marathon**	0.6363	suspect	0.8181	36.57
9.	**Postgresql server dbms database**	0.8867	Data	0.9300	11.99
10.	**Gang rape kangaroo court**	0.7500	Tribal, girl	0.8800	12.44

The table above shows percentage improvement in average relevancy of queries before and after adding new keyword. All the queries improve its relevancy.

V. CONCLUSION

The paper discusses fundamentals of Information Retrieval and Web Information Retrieval. The paper has also dealt with how GA can be used in the field of Web-IR and can efficiently help in web search process for 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 Rogers-Tanimotocoefficient for Web-IR and finding relevant documents has also been discussed. The result was quite

fruitful and GA helped in searching and ranking the documents as par their relevancy with the given query.

REFERENCES

[1] F. G. Erba, Z. Yu, and L. Ting, "Using explicit measures to quantify the potential for personalizing search," ResearchJournal of Information Technology, vol. 3, no. 1, pp. 24–34, 2011.

[2] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, AddisonWesley, New York, NY, USA, 1999.

[3] K. Agbele, H. Nyongesa, and A. Adesina, "ICT and information security perspectives in E-health systems," Journal of Mobile Communication, vol. 4, pp. 17–22, 2010.

[4] J.H.Holland, Adaptation inNatural and Artificial Systems, The University of Michigan Press, Ann Arbor, Mich, USA, 1975.

[5] K. A. DeJong, An Analysis of the Behaviour of a Class of Genetic Adaptive Systems, University of Michigan, 1975.

[6] D. E. Goldberg, Genetic Algorithms in Search, Optimization, Machine Learning, Addison Wesley, 1989.

[7] G. Salton and C. Buckley, "Improving retrieval performance by relevance feedback," Journal of the American Society forInformation Science, vol. 41, no. 4, pp. 288–297, 1990.

[8] L. M. Schmitt, "Fundamental study, theory of genetic algorithms," Theoretical Computer Science, vol. 259, no. 1-2, pp.1–61, 2001.

[9] K. Milena, "Solving timetabling problems using genetic algorithms," in Proceedings of the IEEE 27th InternationalSpring Seminar Electronics Technology: Meeting the Challengesof Electronics Technology Progress, vol. 1, pp. 96–98, 2004.

[10] L. Lin, L. Cao, J. Wang, and C. Zhang, "The applications of genetic algorithms in stock market data mining optimization," in Proceedings of the Capital Market, CRC, Sydney, Australia, 2000.

[11] W. Ying and L. Bin, "Job-shop scheduling using genetic algorithm," in Proceedings of the IEEE International Conferenceon Systems, Man and Cybernetics, pp. 1994–1999, October 1996.

[12] J. F. Frenzel, "Genetic algorithms, a new breed of optimization," IEEE Potentials, vol. 12, pp. 21–24, 1993.

[13] L. Tamine, C. Chrisment, and M. Boughanem, "Multiple query evaluation based on an enhanced genetic algorithm,"Information Processing and Management, vol. 39, no. 2, pp. 215–231, 2003.

[14] M. Koorangi and K. Zamanifar, "A distributed agent based web search using a genetic algorithm," International Journal of Computer Science and Network Security, vol. 7, no. 1, pp. 65– 76, 2007.

[15] R. Varadarajan, V. Hristidis, and T. Li, "Beyond single-page web search results," IEEE Transactions on Knowledge and DataEngineering, vol. 20, no. 3, pp. 411–424, 2008.

[16] S. Maleki-Dizaji, Evolutionary learning multi-agent based information retrieval systems [Ph.D. thesis], Sheffield HallamUniversity, 2003.

[17] J. Cheng,W. Chen, L. Chen, and Y. Ma, "The improvement of genetic algorithm searching performance," in Proceedings of 1st International Conference on Machine Learning and Cybernetics, pp. 947–951, Beijing, China, November 2002.

[18] M. Sinha and S. V. Chande, "Query optimization using genetic algorithms," Research Journal of Information Technology, vol. 2, no. 3, pp. 139–144, 2010.

[19] M. H. Marghny and A. F. Ali, "Web mining based on genetic algorithm," in Proceedings of the AIML O5 Conference, CICC, Cairo, Egypt, December 2005.

[20] S. H. Lin, M. C. Chen, J. M. Ho, and Y. M. Huang, "ACIRD: intelligent Internet document organization and retrieval," IEEE Transactions on Knowledge and Data Engineering, vol. 14, no. 3, pp. 599–614, 2002.

[21] L. C. Chen, C. J. Luh, and C. Jou, "Generating page clippings from web search results using a dynamically terminatedgenetic algorithm," Information Systems, vol. 30, no. 4, pp. 299–316, 2005.

[22] N. Langville and D. Meyer "Science of search engine rankings", Chapter 1 ,Princeton Pubs,2006

[23] B.MPS,A.Kumar " A Primar on the Web Information Retrieval Paradigm", JATIT, March 2007.

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