Journal of Advanced Computing
(2012) 1: 1-8
doi:10.7726/jac.2012.1001
Research Article

Enhanced Search Scheme Precision and Performance using a GA Approach with Application to Arabic Content

Sameh Ghwanmeh*

Received 15 April 2012; Revision received 29 April 2012; Accepted 6 May 2012; Published online 3 June 2012

Abstract

Literature examination shows that information search engines in Arabic are few compared to those available in English and other languages. Additionally, search engines face many problems when programmed in the Arabic language, including difficulty and uncertainty. Employing Genetic Algorithm within the search scheme to improve performance and exactness and tackle issues with non-accurateness of search systems in which Arabic content is used can be considered an advancement. An enhanced search scheme that provides exactness, precision, and performance by applying the Genetic Algorithm Technique to Arabic content is presented in this paper. Based on the user starting page selection, the system employs its dynamic characteristics to search related pages on the Web. A series of experiments has been conducted to test the quality and effectiveness of the proposed system by means of well-known test-base collections – namely, CISI, CACM, and NPL – and 242 Arabic-content sites. General results revealed that the proposed system retrieved the largest number of appropriate documents and minimal non-related documents with respect to user requests in high-performance information retrieval systems that use the Genetic Algorithm.

Keywords: Arabic content; genetic algorithms; web search engines; performance; information retrieval; precision

1. Introduction

Considerable research has been done based on Genetic Algorithm (GA) techniques to improve Web search engines. Results show that GA techniques have the potential to enhance Web search performance and correctness and its application in Information Retrieval (IR). An IR system deals with different data collections. The growing number of Arabic documents on the Web signals the need for advanced and improved Web search engines that retrieve related Arabic documents with high correctness and less time based on user requests. Precision, percentage of the retrieved related-documents and recall are measures used to determine the IR system's effectiveness and correctness (Abdelmgeid, 2007; Hammo, 2009).

*Corresponding author:
Department of Computer Science, Faculty of IT, WISE University, Amman, Jordan
E-mail: sameh@wise.edu.jo

Employing GA in the search engine design provides better performance and correctness of the search results (Aissiou and Guerti, 2009). Theoretically, the GA system's input is represented by a population of individuals called "chromosomes," either randomly generated or created from a pre-knowledge set. The processes of selection, mutation, and crossover in GA system describe the generation of the "chromosome" process. The best of the finest chromosomes represents the GA system output (Lawrence and Giles, 1998; Lawrence and Giles, 1999).

Literature investigation indicates that information search systems available in Arabic are few compared to those programmed in English and other languages. Additionally, search systems face many problems when used with the Arabic language, including difficulty and uncertainty. Also, employing GA algorithm to improve search engine performance and effectiveness and eliminate non-accurateness in Arabic systems is advancement. In this research paper, we have introduced an improved GA-based system that handles problems with Web search engines' low performance, slowness, and non-accurateness. A series of experiments has been conducted to test the performance, quality, and effectiveness of the proposed system by means of the well-known test-base collections CISI, CACM, and NPL, as well as 242 Arabic content sites.

2. Information Retrieval Systems Survey

Much research has been done on conventional IR systems. Abdelmgeid (2007) has presented and analyzed some of these studies as follows:

2.1 Boolean model

A binary index is employed that determines whether a term in one document is significant or not. The query language is used to represent the user's requests based on logical operations that include AND, OR, and NOT. Obviously, the user results from the query processing contain documents that fully match the required criteria (Dean and Henzinger, 1999).

2.2 Vector space mode

A document is represented as a vector in an n-dimensional space (n represents the number of unique terms that can be used to define the documents) and the query is created from the relations delivered in the user request. The IR system retrieves and orders the documents with a greater likeness based on the user request, which means that documents with higher likeness are considered more relevant and consequently must be retrieved by the system with a higher position in the retrieved list (Gibson et al., 1998).

2.3 Probabilistic model

A system's structure and operation mode are totally dependent upon probability theory. Further, Chau and Chen (2008) show that the collection of documents and its related parameters are used to build the parameterized search function criteria.

2.4 Crawler-based search engine

The structure of this type of search engine contains a mechanism to create and update the engine's listings. Google is perhaps the best-known example of a crawler-based search engine. Further examples can be seen in Kluev (2000) and Kumar et al. (1999a).

2.5 Human-powered directories

The construction and its listings depend on humans' submissions and updates. The description submitted is used for search matches. Updating the structure has no meaning for or value to the listing update. Further details and explanations can be seen in Chakrabarti et al. (1999).

3. Web Search Algorithms

There are many popular Web search algorithms undergirding crawler-based engines, such as Breadth-first and Best-first (Kian and Zahedi, 2011). Additionally, Spreading Activation (Chen et al., 1998) and Genetic Algorithm (Cho et al., 1998) techniques are also proposed as a foundation for Web search engines. An additional study dealing with Arabic content application can be found in Aljuaid et al. (2010).

3.1 Breadth-first search

Web pages in the current level are searched in the order in which they were discovered before pages in the next level are searched. Generally, this type of search is used to create document collections for general Web search engines (Kian and Zahedi, 2011).

3.2 Best-first search

This is one of the most common Web search algorithm employed in crawlers. In this algorithm, the heuristics (based on previous search results) are employed in the search ranking and queue order. Non-promising Universal Resource Locators (URLs) are placed in the back of the queue, where they rarely get a chance to be visited (Bergmark, 2002; Bergmark et al., 2002; Chakrabarti et al., 2007). Obviously, this type of search algorithm is more common than the breadth-first search algorithm since it examines the relevant page locations and avoids retrieving non-related pages. Generally, GA executes the evolution process to reach the optimization state by using selection, crossover, and mutation operators (McCallum et al., 1992; Michalewicz, 1996; Gibson et al., 1998; Kumar et al., 1999b).

4. The Proposed Algorithm Description

Many researchers, in the last decade, have considered building Web search engines with high precision and performance. The idea behind the design is the crawling process, in which the requested page or URL is directed to a related Web page (or not) before the fetching process is attained. In general, visiting the requested page is done in an optimal order in which high-quality pages are visited first. The objective of this research is to build a Web search algorithm based on GA that is characterized by higher precision, correctness, and recall compared with conventional Web search algorithms, especially when applied to Arabic content. The proposed algorithm is a combination of conventional Web search and GA techniques. Based on the user starting page selection, the system employs its dynamic characteristics to search the related pages on the Web. The proposed GA algorithm is presented in Fig. 1. Additionally, the GA uses a crossover operation, which involves the mixing of two chromosomes to create a new one, as well as a mutation operation, which includes the adjustment of the gene values of a result with a given probability to yield the required chromosomes. The performance measure employed in this GA is the cosine fitness function adopted by Abdelmgied (2007). Further details about these GA operators and the fitness function can be found in Abdelmgeid (2007).

Step #1: Initial solutions (samples) are randomly generated.
Step #2: All solutions' (samples) fitness function is measured. Reproduction and selection process are initiated based on fitness function by using different selection methods, such as stochastic universal and rank tournament.
Step #3: The selected individuals are processed using the GA operators (inversion, crossover, and mutations) to generate new ones. These GA operators allow to reach good solutions and to create new solutions.
Step #4: Repeat Step #2 until the ending state is reached.

Figure 1: The proposed Genetic Algorithm applied to Arabic content

5. Experimental Results and Discussion

A series of experiments has been conducted to test the proposed algorithm's performance, correctness, and quality. After the system is initialized with the initial parameters and relevant values to reach the current generation set, then the user inputs the simple query Web page URLs and the search keywords, such as information retrieval, input starting URLs, and keywords. The universal keyword sets (S1–S5) are built by expanding the relevancy of each keyword, as shown in the example below. The keyword domain chromosome is created for each initial URL. The chromosome structure consists of a series of 1's and 0's. 1 means the keyword is in the S, while 0 means it is not in the S. Table 1 shows the initial GA algorithm form of the chromosome. Alternatively, user requests and user rankings of collective sets of keywords are used to form the most appropriate keyword domain. Table 2 presents the chromosome structure for both fixed and dynamic keyword domains. The relevance of particular keywords, based on existing domain, is represented on a standard scale (0 to 10) that is presented in Table 2, while Table 3 presents the URL and its ranking values. There two options to establish the value of the chromosome set, either by using the standard scale (Table 2) or by using the chromosomes' dynamic size (Table 3).

5.1 Example of universal keyword sets

- S1: Computer database, IR, Computer networks, "شبكات الحاسوب," improvements, approach, multiple, query, relation, " علاقة," relational, retrieval, Databases, "قواعد البيانات," queries, relational databases, "قاعدة بيانات علائقية," relational database, us, carat.dat, gqp.dat, orus.dat, query. Opt
- S2: information, "معلومات," information retrieval, information storage, "وعاء معلومات," indexing, retrieval, "تنقيب," storage, us.
- S3: Artificial intelligence, "ذكاء صناعي," information retrieval systems, "نظم التقيب عن البيانات," information retrieval, indexing, natural language processing, "معالجة اللغات الطبيعية," us, dbms.ai.
- S4: Fuzzy set theory, "النظرية الضبابية," information retrieval systems, indexing, performance, "كفاءة," retrieval systems, "نظم التنقيب عن البيانات," retrieval, queries, us.
- S5: Information retrieval systems, "نظم التنقيب عن البيانات," indexing, "الفهرسة," retrieval, stairs, us.

5.2 Collective set of all keywords

Computer database, IR, Computer networks, "شبكات الحاسوب," improvements, approach, multiple, query, "علاقة," relation, relational, retrieval, Databases, "قواعد البيانات," queries, relational databases, "قاعدة بيانات علائقية," relational database, us, carat.dat, gqp.dat, orus.dat, query. Opt, information, "معلومات," information retrieval, information storage, "وعاء معلومات," indexing, retrieval, "تنقيب,"

storage, us, Artificial intelligence, "ذكاء صناعي," information retrieval systems, "نظم التقيب عن البيانات," information retrieval, indexing, natural language processing, "معالجة اللغات الطبيعية," us, dbms.ai, Fuzzy set theory, "النظرية الضبابية," information retrieval systems, indexing, performance, "كفاءة," retrieval systems, "نظم التنقيب عن البيانات," retrieval, queries, us, Information retrieval systems, " نظم التنقيب عن البيانات," indexing, "الفهرسة," retrieval, stairs, us.

Table 1: Initial GA form of chromosome

Genetic pattern of chromosome	Chromosome fitness
11111111111111111000000000000000	[0.267434]
00001000000100010000111110000010	[0.217652]
00011000011000010100001001111000	[0.457578]
00000011000110010000010101001110	[0.325263]
00001100000100010001001010100010	[0.524132]

Table 2: The chromosome structure for fixed and dynamic keyword domain

high in	in		do not know		not in	
10		8	4		0	
k1	k2	k3	k4	k5	k6	k7
10	0	10	4	8	4	0

Table 3: URL and its ranking value

Set (URL)	Ranking Value				
S1	k1	k4	k5	k6	
S2	k2	k3	k6		
S3	k1	k2	k4	k6	k7

5.3 Content-based analysis selection

The Jaccard's similarity function is used through this selection method. The fitness function value of the objective Web page is calculated, which represents the likeness between the page and a domain dictionary. Larger fitness values represent more pages that are related to the domain dictionary and, hence, are more likely to be appropriate to the target domain. Repeated process is applied to calculate all fitness values of all S's and URLs in the current generation. The resulted pages with higher fitness values are sorted and the best stemmed chromosome, which is structured based on the optimal keyword domain, will be submitted by the user to keep a replica for the purpose of alteration operations and continue the same process. Otherwise, the previous process will be repeated until the optimal keyword domain is produced according to user requirements. The larger fitness values determine the probability of a page to persist in this selection procedure by using a contest selection. Subsequently, the persisting pages are retained in local page depositories and the residual pages are marked as irrelevant (Table 4, Table 5, and Table 6). Results revealed that the average fitness (Jaccard's score) of the starting URL is 0.27643 and the resulting optimized chromosomes in the population are given in Table 7. As most of the previous studies of the search schemes deal with collection of English content only, a comparison with a conventional search

scheme is not appropriate. However, it can be seen that the proposed GA scheme yields better search results compared to the conventional and non-Arabic IR system presented in Abdelmgeid, (2007). The GA scheme has a direct effect on the number of terms and the fitness function value.

Table 4: Fixed or dynamic resultant keyword domain chromosome (S1 and S2)

	k1	k2	k3	k4	k5	k6	k7
S1 (URL)	1	0	0	1	1	1	0
S2 (URL)	0	1	1	0	0	1	0

Table 5: Fixed or dynamic resultant keyword domain chromosome (S1 and S3)

	k1	k2	k3	k4	k5	k6	k7
S1 (URL)	1	0	0	1	1	1	0
S3 (URL)	1	1	0	1	0	1	1

Table 6: Jaccard results

Set	Score
S1 & S1	1.000
S1 & S2	0.130
S1 & S3	0.130
S1 & S4	0.1143
S1 & S5	0.00

Table 7: Resultant optimized chromosomes

Genetic pattern of chromosome	Chromosome fitness
00000000000100010000010101000001	[0.4310]
00000000000100010000010101000001	[0.4310]
00000000000100010000010101000001	[0.4310]
00000000000100010000010101000001	[0.4310]

6. Conclusion

This paper presented an enhanced GA-based system that handles the problems of Arabic Web search engines' low performance, slowness, and non-accurateness and conducted a series of experiments to test the performance, quality, and effectiveness of the proposed system by means of well-known test-base collections CISI, CACM, and NPL as well as 242 Arabic content sites. The proposed technique is used to determine domain collections for standard search engines. The fitness function is calculated based on the user query. Additionally, a mechanism during the initialization phase is presented to the user according to his information about the required query and also provides a real presentation of the chromosomes used in the earlier techniques. The proposed GA algorithm yields a set of Web pages and is targeted at maximizing the fitness function. The resulting pages with higher fitness values are sorted and the best stemmed chromosome, which

is structured based on the optimal keyword domain, is submitted by the user to keep a replica for alteration purposes and to continue the same process. Results revealed that the average fitness (Jaccard's score) of the starting URL is 0.27643 and that the proposed system provided optimized chromosomes in the population when applied to the standard test collection, including Arabic content.

Acknowledgement

The researcher gratefully acknowledges and highly appreciates the financial support and the remarkable resources provided by WISE University, Amman, Jordan.

References

Aljuaid, H., Zulkifli, M., Sarfraz, M., 2010. A tool to develop Arabic handwriting recognition system using genetic approach. Journal of Computer Science 6(6), 619-624.

Aissiou, M., Guerti, M., 2008. Genetic algorithm application to the standard Arabic phonemes classification. Cybernetics & Systems 39(3), 199-212.

Abdelmgeid, A., 2007. Applying genetic algorithm in query improvement problem. Int. J. Inform. Technol. Knowl. 11, 309-316.

Bergmark, D., 2002. Collection synthesis. Proceeding of the Joint Conference on Digital Libraries, Portland, Oregon, USA, 253-262.

Bergmark, D., Lagoze, C., Sbityakov, A., 2002. Focused crawls, tunneling, and digital libraries. Lecturer Notes Comput. Sci. 2458, 91-106. http://dx.doi.org/10.1007/3-540-45747-X_7

Chau, M., Chen, H., 2008. Comparison of three vertical search spiders. IEEE Comput. 36, 56-62. http://dx.doi.org/10.1109/MC.2003.1198237

Chakrabarti, S., van den Berg, M., Dom, B., 1999. Focused crawling: A new approach to topic-specific Web resource discovery. Comput. Networks: Int. J. Comput. Telecommun. Network. 31, 1623-1640.

Chakrabarti, S., B. Dom, S., Kumar, R., Raghavan, P., Rajgopalan,S., et al., 2007. Mining the Web's link structure. IEEE Comput. 32, 60-67.

Chen, H., Chung, Y., Ramsey, M., Yang, C., 1998. A smart itsy-bitsy spider for the Web. J. Am. Soc. Inform. Sci. 49, 604-618. http://dx.doi.org/10.1002/(SICI)1097-4571(19980515)49:7<604::AID-ASI3>3.0.CO;2-T

Cho, J., Garcia-Molina, H., Page, L., 1998. Efficient crawling through URL ordering. Proceeding of the 7th International World Wide Web Conference, April 14-18, Brisbane, Australia.

Cho, J. and Garcia-Molina, H. and Page, L. (1998) Efficient Crawling Through URL Ordering. In: Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia

Dean, J. and Henzinger, M., 1999. Finding related pages in the World Wide Web. Proceeding of the 8th International WWW Conference, Apr. 14-18, Toronto, Canada, 1-20. http://ilpubs.stanford.edu:8090/347/1/1998-51.pdf.

Gibson, D., Kleinberg, J., Raghavan, P., 1998. Inferring Web communities from link topology. Proceeding of the 9th ACM Conference on Hypertext and Hypermedia, June 20-24, ACM Press, Pittsburgh, Pennsylvania, USA, 225-234. http://portal.acm.org/citation.cfm?id=276652.

Hammo, B., 2009. Towards enhancing retrieval effectiveness of search engines for diacriticized Arabic documents. Information Retrieval 12(3), 300-323. http://dx.doi.org/10.1007/s10791-008-9081-9

Kian, H. and Zahedi, M., 2011. An efficient approach for keyword selection: Improving the accessibility of Web contents by general search engines. International Journal of Web & Semantic Technology 2(4), 81-90. http://dx.doi.org/10.5121/ijwest.2011.2406 PMid:16735669

Kluev, V., 2000. Compiling document collections from the Internet. SIGIR Forum 34, 9-14.

http://dx.doi.org/10.1145/381258.381264
Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A., 1999a. Trawling the Web for emerging cyber-communities. Proceeding of 8th International World Wide Web Conference, Toronto, Canada, 1-13. http://www.uzh.ch/home/mazzo/reports/www8conf/2166/pdf/pd1.pdf.
Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A., 1999b. Extracting large-scale knowledge bases from the Web. Proceeding of the 25th International Conference on Very Large Data Bases conference, Edinburgh, Scotland, UK, 1-12. http://www.vldb.org/conf/1999/P60.pdf.
Lawrence, S., Giles, C. Lee, 1998. Searching the World Wide Web. Science 280, 98-100. http://dx.doi.org/10.1126/science.280.5360.98 PMid:9525866
Lawrence, S., Giles, C. Lee, 1999. Accessibility of information on the Web. Nature 400, 107-109. http://dx.doi.org/10.1038/21987 PMid:10428673
Michalewicz, Z., 1996. Genetic algorithms + data structures = Evolution Programs. Springer-Verlag, Heidelberg, ISBN: 10: 3540606769.
McCallum, A., Nigam, K., Rennie, J., Seymore, K., 1992. Building domain-specific search engines with machine learning techniques. Proceeding of the AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace, Orlando, Florida, USA, 1-6.