ARTICLE

JOURNAL OF UNIVERSITY OF BABYLON
For Pure and Applied Sciences (JUBPAS)

Vol.30; No.3.| 2022

Adaptive Indexing of Documents Using Genetic Algorithms and Relevance Feedback

Tefool H. O. Al-khafaji [1] and Ali Hasan Shaheed [2]

1 College of Education for Pure Sciences, University Of Babylon, pure.tefool.hussein@uobabylon.edu.iq, Babylon, Iraq

2 College of Science for Women, University Of Babylon, wsci.ali.hasan@uobabylon.edu.iq, Babylon – Iraq

***Corresponding author email:** engineer.ali.alkhafaji2015@gmail.com; mobile:07829490175

الفهرسة التكيفية للوثائق باستخدام الخوارزميات الجينية والتغذية الراجعة ذات الصلة

طفول حسين عمران الخفاجي[1]، علي حسن شهيد [2]

1 كلية التربية للعلوم الصرفة، جامعة بابل، pure.tefool.hussein@uobabylon.edu.iq ، بابل، العراق.

2 كلية العلوم للبنات، جامعة بابل، wsci.ali.hasan@uobabylon.edu.iq ، بابل، العراق.

Received:	1/4 /2022	Accepted:	23 /6 /2022	Published:	30 /9 /2022

ABSTRACT

Background:

In this paper, the problem of retrieving the correct documents that satisfy the user's concerns is investigated. The main aim in information retrieval systems is to retrieve all and only relevant documents.

Materials and Methods:

The genetic algorithm is utilized to adapt and change the documents indexes, depending on relevance judgments collected from users. Genetic algorithm is a powerful tool that depends on the Darwinian principles and evolution techniques to search complex spaces. The use of genetic algorithm facilitates the adaptation of documents indexes. Sampling operation is performed using roulette wheel, roulette wheel with elitism and stochastic universal sampling. The fitness function is computed using Jaccard's coefficient that measure the closeness between query and document index.

Results:

The results show that the new descriptions are more efficient and closer to the population of users that use the information retrieval system. In addition, the stochastic universal sampling gave the best results.

Conclusion:

The keywords used to describe the content of documents have statistical dependencies among them. It is difficult to accommodate these dependencies in retrieval system. Genetic algorithm can consider these dependencies during its action. According to ***schema theorem*** and ***building block hypothesis*** [10], the fittest schemata are propagated from generation to generation, where they are sampled, recombined, mutated and resampled to form strings of potentially higher worth. Another aspect genetic algorithm can offer, is the reliance on the feedback provided by users of the retrieval system to adapt documents descriptions and selections variations were experimented with roulette sampling, with elitism, and with produce new set of descriptions closer to the population of users' needs.

Three fitness proportionate selection variations are used, roulette wheel sampling, roulette wheel with elitism and stochastic universal sampling. The results have indicated the superiority of the third over the first two.

Key words:

Adaptive Indexing, relevance feedback, document retrieval, genetic algorithm

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

الخلاصة

مقدمة:

في هذه الورقة، تم البحث في مشكلة استرجاع الوثائق الصحيحة التي تحقق رغبات المستخدم. الهدف الرئيسي في أنظمة استرجاع المعلومات هو استرداد جميع الوثائق ذات الصلة فقط.

طرق العمل:

تم استخدام الخوارزمية الجينية لتحقيق هذا الهدف. أوصاف المستندات تم تكييفها وتغييرها باستخدام الخوارزمية الجينية، اعتمادًا على الأحكام التي اطلقها المستخدم (والتي تم جمعها والاحتفاظ بها) حول اهمية المستندات بالنسبة له. الخوارزمية الجينية هي أداة قوية تعتمد على مبادئ الداروينية وتقنيات التطور للبحث في فضاءات البحث المعقدة. يسهل استخدام الخوارزمية الجينية تكييف فهارس المستندات. تم تنفيذ ثلاثة طرق في الانتخاب: نمذجة عجلة الروليت ، ونمذجة عجلة الروليت مع النخبة والنمذجة الشاملة التصادفية. يتم حساب دالة الصلاحية باستخدام معامل Jaccard الذي يقيس التقارب بين الاستعلام وفهرس المستند.

الإستنتاجات:

توجد بين الكلمات المفتاحية المستخدمة لوصف محتوى الوثائق اعتماديات إحصائية. من الصعب استيعاب هذه الاعتماديات في نظام الاسترجاع. يمكن للخوارزمية الجينية أن تأخذ في الاعتبار هذه الاعتماديات أثناء عملها. وفقًا لنظرية المخطط وفرضية حجر البناء [10] ، يتم نشر المخططات الأكثر صلاحية من جيل إلى جيل ، حيث يتم أخذ عينات منها وإعادة تجميعها وتحويلها وإعادة تشكيلها لتشكيل سلاسل ذات صلاحية أعلى. هناك جانب آخر يمكن أن تقدمه الخوارزمية الجينية ، وهو الاعتماد على التغذية الراجعة المقدمة من مستخدمي نظام الاسترجاع لتكييف أوصاف المستندات، وإنتاج مجموعة جديدة من الأوصاف الأقرب إلى حاجات المستخدمين. تم استخدام ثلاثة انواع من الانتخاب المتناسب مع الصلاحية ، وهي نمذجة عجلة الروليت ، ونمذجة عجلة الروليت ذات النخبة ، والنمذجة الشاملة التصادفية. أظهرت النتائج تفوق النوع الثالث على الأول والثاني.

الكلمات المفتاحية:

النمذجة التكيفية، التغذية الراجعة ذات الصلة، استرجاع الوثائق، الخوارزمية الجينية.

INTRODUCTION

The problem of retrieving documents is an active research area because of the increasing need to retrieve information from huge databases [1]. The main goal of any information retrieval system is to retrieve all (recall) and only (precision) relevant documents. Many researches have been done to achieve this goal. Some of them that used genetic algorithms will be reviewed in this section. Pathak et.al [2] investigated the possibility of applying genetic algorithms to adapt various matching functions. An overall matching function is treated as a weighted combination of scores produced by individual matching functions. This overall score was used to rank and retrieve documents. Weights associated with individual functions were searched using Genetic Algorithm. The idea was tested on a real document collection and the results were promising. Horng and yeh [3] produced an approach to automatically retrieve keywords and then genetic algorithm was used to adapt the keywords weights. Radwan et-al. [4] presented a new fitness function for approximate information retrieval and they found that it was very fast and very flexible than cosine similarity fitness function. Abualigah et-al [5] used cosine similarity and Jaccards to compute similarity between the query and documents, and used two proposed adaptive fitness function, mutation

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

operators as well as adaptive crossover. The process aimed at evaluating the effectiveness of results according to the measures of precision and recall. Abdul Hassan et-al and, Kulunchakov and Strijov [6 and 7], new matching functions are proposed to enhance retrieval of documents in terms of recall and precision in Abdul Hassan et-al [6], and in terms of solving the problems of stagnation and complexity in Kulunchakov and Strijov [7]. Vasgi and Kulkarni [8] worked in secured environments and employed genetic algorithm as a matching tool to measure similarities between queries and documents. It aimed to increase the relevance in secure encrypted domain and find the best effective search technique over cloud data with enhanced security. Irfan and Kumar [9] used the genetic algorithm as a ranking tool to rank the results of user's query. The approach they propose helps in optimizing the retrieved information in minimum time duration by following a limited set of iterations.

Materials and Methods

- **Genetic Algorithms**

Genetic algorithms are search and optimization tools that inspired by biological evolution and Darwinian principles of survival of the fittest and natural selection. J.H. Holland proposed genetic algorithm in 1975. Genetic Algorithms work on a population of chromosomes, each of which represents a solution to the problem to be solved. These chromosomes are transformed using a coding scheme into a form suitable to a genetic algorithm. The fitness function allocates values to all chromosomes in the population to determine their efficiency. The population of chromosomes is processed using genetic operators: selection, crossover and mutation. The use of genetic operators leads to the replacement of the current generation with a new, more efficient generation. In selection, the chromosomes are selected based on its fitness value for further processing. In crossover operator, a random locus is chosen and it changes the subsequences between chromosomes to create off-springs. In mutation, some bits of the chromosomes will be randomly flipped based on probability. [10] [11]. In this work, genetic algorithm is employed to improve the information retrieval system. The proposed system is utilizing feedback from users of the system to re-describe documents in order to satisfy inquirers needs.

- **Adaptive Indexing of Documents Using Genetic Algorithms**

The proposed system is composed of two subsystems: Document Retrieval System (DRS) and Genetic Adaptation System (GAS). During the operation of DRS, the queries and relevance judgements are collected and saved in tables. This saved information is used by the GAS to change the description of documents.

In this work, the retrieval process is based on the vector space model. In such a model, documents can be viewed as binary vectors whose components indicate the presence or absence of the ***i***th indexing term.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

- **Document Retrieval System (DRS)**

The proposed system employed binary vectors to describe the content of the documents as well as queries. The binary vector has the form:

$$D_i = < \begin{matrix} T_1, & T_2, & \dots & T_m \\ d_{i1} & d_{i2} & \dots & d_{im} \end{matrix} >$$

Where $\mathbf{T_1, \dots, T_m}$ are index terms (or Phrases) such that the term $\mathbf{T_j}$ is either used to index a document ($\mathbf{d_{ij}=1}$) or not ($\mathbf{d_{ij}=0}$). The Jaccard's coefficient [1] is used as a measure to estimate association between query and document. The computation of relevance is accomplished as follows:

{*Jaccard's Coefficient*}

Query: $Q=<q_1, q_2, \dots, q_m>$, Where q_i is the value of the ith query term (0,1).

Document: $D_j=<d_{j1}, d_{j2}, \dots, d_{jm}>$, where d_{ji} is the value of the ith term in document$_j$.

$$\text{Similarity Function } S(Q, D_j) = \frac{\sum_{i=1}^{m}(q_i . d_{ji})}{\sum_{i=1}^{m}\left(max(q_i, d_{ji})\right)} \quad 0 \le S \le 1$$

Two main steps are performed in the matching subsystem: *matching*, and *ordering and retrieving*. In the matching step, the query and documents' descriptions are converted from a set of keywords to binary vectors, then the query is matched serially against all the documents stored in the database, that contains current query's search terms, and relevance values are computed. The second step, ordering and retrieving, reorders documents according to the relevance values that were computed in the previous step. The retrieved documents will have relevance values in the range:

$$0 < relevance \le 1$$

Where 1 is the maximum relevance that may be obtained and the value 0 is a threshold given to the matching function.

The retrieved documents are displayed for the user to browse them and give his relevance judgments on each retrieved document. The relevance judgments are important for the GAS to perform its operation of adapting documents' descriptions. After collecting sufficient relevant and non-relevant queries, the second system, GAS, will begin its operation of re-describing documents depending on the queries and relevance judgments of population of users who used the system. This means, the experience of the users is utilized to influence documents' descriptions to yield

better levels of performance for future inquirers. A group of ***Related Queries*** and ***Non-Related Queries*** are collected for each document studied.

- **Genetic Adaptation System (GAS)**

Refine document index means to repeat the indexing process to represent it optimally for the set of users who will find it beneficial. The adaptation model is this: a document is represented by different full indices; a user produces a query; and each index of the document is compared with the query as if the document was indexed with only an individual index. The average of these scores is calculated. These separate matching scores and their average serve as feedback information that the genetic algorithm exploits to re-index the document.

A way to adapt document index is to gather a set of users' queries that judged to be related to the document and a set of queries that judged to be non-related to the document. This operation is made for each document studied. This means, the experience of the users is utilized to influence documents descriptions to yield better levels of performance for future inquirers. Each document studied was dealt with independently; i.-e. each document had its own set of indices and its own set of related and non-related queries.

After collecting sufficient number of relevant and non-relevant queries, the GAS starts its operation to improve the descriptions of documents. The genetic algorithm is the adaptation tool that used to enhance the descriptions of documents using feedback information collected previously.

The features of the genetic algorithm that is used in this work is indicated in Table 1. The genetic algorithm main loop to re-describe documents is indicated in figure 1.

{Genetic Algorithm: Main Loop}

1. Open documents Table.
2. Get first document.
3. **While** not end of document table **do**:
    a. Initiate descriptions, related queries and non-related queries for current document.
    b. Calculate fitness for each description depending on its resemblance to related queries and its difference from non-related queries.
    c. **While** GA termination criterion not satisfied **do**:
        i. Apply genetic operators on descriptions.
    d. Calculate fitness for each description depending on its resemblance to related queries and its difference from non-related queries.
    e. Get next document.
4. Discard original documents descriptions and index documents with new promoted descriptions.

Figure 1: Genetic Algorithm Main Loop.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Results and Discussion

The genetic algorithm was applied to the task of document re-description and its effectiveness was tested experimentally. A set of thirty documents was used, each of which has eighteen relevant queries, eighteen non-relevant queries and eighteen descriptions. The genetic algorithm was run for 50 generations. The number of generations was the condition for the algorithm.

Three variations were experimented:

1. Roulette wheel sampling (RW).
2. Roulette wheel sampling and elitism.
3. Stochastic universal sampling (SUS).

Table 1. Genetic Algorithm Features

Genetic Algorithm Features	
Initial Population	The set of all relevant queries collected previously.
Fitness Function	$\boldsymbol{RF + wt \times NF}$ where: RF: relevance fitness estimated using Jaccard's coeffiecient. NF: non-relevance fitness estimated using Jaccard's coeffiecient. wt.: weight (experimentally found wt.=0.7)
Selection	Fitness Proportionate Selection
Sampling	• Roulette Wheel Sampling • Roulette Wheel Sampling with elitism • Stochastic Universal Sampling
Crossover	One-point crossover Crossover probability = 1
Mutation	Simple genes flip Mutation Probability = 0.001

Fitness proportionate selection with SUS gave better results than other two variations experimented. It is the best in terms of the best individuals detected during GA operation. Roulette

wheel sampling was the worst one. With 'elitism', better solutions are obtained which were very close to the SUS results (see table 2). In table 2, only 10 documents are presented.

Table 3 indicates the document retrieval system responses to some queries before and after adaptation. We notice that, most of the queries' relevance to the descriptions of documents was increased after adaptation, such as query 1 with documents 1 and 4. Another set of queries' relevance to the descriptions of documents were decreased, such as query 7 with document 3. We see that an entirely new set of keywords can be embodied in indexing a document. The term "inquirer indeterminacy" was not utilized in the initial description of document 7. After adaptation, it was not incorporated in document description. On the other hand, the term "natural scene images" was employed in the initial set of descriptions to document 19 while it is not after adaptation.

Table 2. Fitness Improvement to three variations Experimented

	Fitness proportionate selection with RW			Fitness proportionate selection with elitism		Fitness proportionate selection with SUS	
	Gen 1	Gen 50	Change %	Gen 50	Change %	Gen 50	Change %
Doc 1	19.4	27.85	43.56	30.29	56.13	33.19	71.08
Doc 2	14.3	21.4	49.65	23.21	62.31	23.54	64.62
Doc 3	18.54	26.68	43.91	26.31	41.91	26.15	41.05
Doc 4	22.65	31.91	40.88	29.96	32.27	31.46	38.9
Doc 5	21.34	26.37	23.57	28.47	33.41	29.71	39.22
Doc 6	18.46	29.66	60.67	30.41	64.73	31.82	72.37
Doc 7	18.29	27.74	51.67	27.45	50.08	28.41	55.33
Doc 8	18.35	22.67	23.54	27.43	49.48	29.21	59.18
Doc 9	17.75	28.92	62.93	28.64	61.35	29.88	68.34
Doc 10	18.8	26.04	38.51	26.34	40.1	29.27	55.69

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Table 3. Document Retrieval System Before and After Adaptation

Qn	Query	Before Adaptation		After Adaptation	
		Doc. No.	Relevance	Doc. No.	Relevance
1	Genetic Algorithm	4 1	0.125 0.08	4 1	0.25 0.2
2	Term discrimination Term precision	13 2	0.33 0.17	13	1
3	Term discrimination	13 2	0.2 0.17	13	0.5
4	Inquirer inconsistency	None		7	0.4
5	Inter-indexer inconsistency, Inquirer inconsistency	7	0.14	7	0.4
6	Adaptive retrieval, Learning	4 11	0.11 0.08	1 4	0.4 0.2
7	Relational DBMS, SQL	3	0.4	3	0.29
8	Conjunctive normal form, ad hoc queries	3	0.17	3	0.125
9	Conjunctive normal form	None		3	0.14
10	Automatic query processing, query formulation	None		3	0.125

Conclusion

The keywords used to describe the content of documents have statistical dependencies among them. It is difficult to accommodate these dependencies in retrieval system. Genetic algorithm can consider these dependencies during its action. According to ***schema theorem*** and ***building block hypothesis*** [10], the fittest schemata are propagated from generation to generation, where they are sampled, recombined, mutated and resampled to form strings of potentially higher worth. Another aspect genetic algorithm can offer, is the reliance on the feedback provided by users of the retrieval system to adapt documents descriptions and produce new set of descriptions closer to the population of users' needs.

Three fitness proportionate selections variations were experimented with roulette sampling, with elitism, and with stochastic universal sampling. The results have indicated the superiority of the third over the first two.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Conflict of interests.

There are non-conflicts of interest.

References

[1] Christopher D. Manning, Prabhakar Ravaghan, Hinrich Schutze, An Introduction to Information Retrieval, Cambridge University Press, 2009.

[2] Praveen Pathak, Michael Gordon, Weiguo Fan, Effective Information Retrieval using Genetic Algorithms based Matching Functions Adaptation, Proceedings of the 33rd Hawaii International Conference on System Sciences, 2000.

[3] Jorng-Tzong Horng, Ching-Chang Yeh, Applying genetic algorithms to query optimization in document retrieval, Information Processing and Management 36, pp. 737-759, 2000.

[4] Ahmed A. A. Radwan, Bahgat A. Abdel Latef, Abdel Mgeid A. Ali, and Osman A. Sadek, Using Genetic Algorithm to Improve Information Retrieval Systems, International Journal of Computer and Information Engineering Vol.2, No.5, 2008.

[5] Laith Mohammad, Qasim Abualigah, Essam S. Hanandeh, Applying Genetic Algorithms To Information Retrieval Using Vector Space Model, International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.5, No.1, 2015.

[6] Alia Karim Abdul Hassan, Duaa Enteesha mhawi, Sarah Najm Abdulwahid, Information Retrieval using Modified Genetic Algorithm, Al-Mansour Journal, 27, 2017.

[7] A.S. Kulunchakov, V.V. Strijov, Generation of simple structured information retrieval functions by genetic algorithm without stagnation, Expert Systems With Applications, 85, pp. 221-230, 2017.

[8] Bharati P. Vasgi, Uday V. Kulkarni, Secure Retrieval in Outsourced Environment using Genetic Algorithm, 5th International Conference on Computing Communication Control and Automation (ICCUBEA), 2019.

[9] Shadab Irfan, D. Rajesh Kumar, Ranking Algorithm to Optimize the Retrieval Process Using Genetic Algorithm, International Journal of Control and Automation Vol. 13, No. 2, pp. 383 – 396, 2020.

[10] David E. Goldberg, Genetic Algorithms in Search Optimization and Machine Learning, Addison-Wesley, 1989.

[11] Sourabh Katoch, Sumit Singh Chauhan and Vijay Kumar, A review on genetic algorithm: past, present, and future, Multimedia Tools and Applications, 80, pp. 8091–8126, 2021.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com