V.K. Ivanov, B.V. Palyukh, A.N. Sotnikov

EFFICIENCY OF GENETIC ALGORITHM FOR SUBJECT SEARCH QUERIES

Tver State Technical University, Tver State Technical University, Joint Supercomputer Centre of the Russian Academy of Sciences

E-mail address: mtivk@tstu.tver.ru, pboris@tstu.tver.ru, asotnikov@jscc.ru

Received 2015/11/09

Abstract. The article presents and generalizes the results on some performance indicators of genetic algorithm developed by authors and applied to effective search queries and selection of relevant results after document subject search. It is shown that the developed technology expands opportunities of semantic search and increases the number of the found relevant results. In particular, we made an effort to show the ability of the developed algorithm to achieve the neighborhood of the fitness function in a finite number of steps, to provide higher precision of search in comparison with the well-known search engines of the Internet as well as to provide the acceptable semantic relevance of the found documents.

1. Introduction

The subject search in document storages [1], [2], [3] is a well-established procedure. Experience has shown, however, that its efficiency continues to be a challenge which is not simple as it may seem.

2010 Mathematical Subject Classification. 68W40.

Key words and phrases. convergence, genetic algorithm, fitness function, population, ranking, relevance, search precision, search query.

This work was supported by the Russian Foundation of Basic Research (contract No 13-07-00342). The authors express their thanks the colleagues who have been taking part in this study on an expert basis, namely, A.A. Gusarov, I.A. Egereva, D.A. Zorenko, D.G. Kuzmin, I.V. Obraztsov, A.P. Panasenkov V.B. Petropavlovskaya, Yu.P. Rumyantsev, G.P. Shmatov as well as to translators of the article V.V. Sizova and O.A. Gumenyuk. .

The subject document search is aimed at selecting documents with coordinated information (interrelated facts, their retrospection and perspective) in a subject segment or of a specified object. The search brings a great number of documents being maximally relevant to the subject area on the whole and not just information about single events, objects or phenomena. The subject search areas are as follows: search for innovation solutions, development of new business opportunities, clients information collection, competitive analysis and intelligence, scientific and technical information reviews, project examinations, patent research, teaching material selection. Efficient search algorithms play a key role in subject search since the users inevitably face the following objective problems:

• the difficulty of selecting coordinated key notions for wording the query;

• the query composition, structure, and complexity limitations of the retrieval system;

• the fragmentation and heterogeneity of target information, availability of alternatives with compatible relevance;

• the absence of effective search result clustering/classification systems.

Solving the stated problems one should properly interpret search results. This means the simultaneous relevance estimation of documents found with different queries, retrieval system ranking perfectness, availability of all relevant results for estimation, availability of effective solutions in other areas for successful use in this area.

One of the main difficulties here is a semantically correct statement of search queries to provide acceptable measures of search efficiency. The article presents and generalizes the study performance results of the genetic algorithm used for generating efficient search queries and selecting relevant search results.

The research has become the extension of the works performed by authors before, and included:

• assessment of the worked out genetic algorithm convergence. We made an effort to show the ability of the developed algorithm to achieve the neighborhood of the fitness function optimum (i.e. to reveal and select the most relevant documents) in a finite number of steps;

• assessment of the search results relevance. We showed the ability of the developed genetic algorithm to provide higher precision of search (i.e. share of relevant documents in the search results) in comparison with the well-known search engines of the Internet.

Here, take into account the fact that high precision of search of the worked out algorithm provides the acceptable semantic relevance (pertinence) of the found documents. Note that the detailed description of calculation method of the objective function of genetic algorithm presented in the article will help you to understand our approach to the assessment of document relevance. We would like to emphasize that all simulation experiments were carried out by means of the developed algorithm program implementation.

2. Related Works

This is not to say that we were the first to use genetic algorithms to generate search queries. In general, evolutionary algorithms are quite intensively used in solving the information retrieval problems including search and analysis of data in the Web (see, e.g. reviews [4] and [5]). It is notable that there are at least two trends in the development of models based on genetic algorithms. The first of them interpret the populations of different generations as sets of individuals, i.e. web-pages found by search engine. The result of this evolutionary process is a ranked list of web-pages which is more relevant than suggested by the search engine.

For instance, [6] develops the page clipping synthesis (PCS) search method to extract relevant paragraphs from other web search results. The PCS search method applies a dynamically terminated genetic algorithm to generate a set of best-of-run page clippings in a controlled period of time. Effectiveness measure confirmed that PCS performs better than standard search engines.

[7] suggests using the evolutionary techniques to derive good evidence combination functions by using three different sources of relevance evidence: the textual content of documents, the reputation of documents extracted from the connectivity information available in the processed collection and the anchor text concatenation. The experiments performed indicate that this proposal is an effective and practical alternative for combining sources of evidence into a single ranking.

The paper [8] proposes an approach for web content mining using genetic algorithm. The proposed approach considers several parameters like time website existed, backward and forward links. It has been shown experimentally that this approach is able to select good quality web pages as compared to the other existing algorithms.

The authors of [9] propose a genetic search for search engines. They show that there is an important relation between web statistical studies and search engines standard techniques in optimization. They put forward a fitness function to assess the quality of the executed crossings and mutation. The paper presents the results of experimental studies by which the quality of final pages is significantly better than that of standard search engines.

The second trend is related to the optimization of queries used by a search engine to obtain the relevant list of required documents. Here populations are sets of individuals, i.e. context-sensitive search queries generated from a given set of key words and concepts. In this case, fitness function evaluates the quality of results of each query individually and the population of queries taken as a whole.

For example, [10] gives a genetic approach that combines the results of multiple query evaluations. The genetic algorithm aims at optimization of the overall relevance estimate by exploring different directions of the document space.

[11] and [12] discuss optimization techniques based on genetic algorithms to evolve good query terms in the context of a given field. The proposed technique place emphasis on searching for novel material that is related to the search context. The authors suggest applying a mutation pool to allow the generation of queries with new terms, prove the effectiveness of different mutation rates on the exploration of query-space. The developed fitness function, which favors the construction of queries containing novel but related terms, is also of our interest.

[13] offers a technique that gives a keyword query, generates new pages called composed pages containing all query keywords. The composed pages are generated by extracting and stitching together the relevant pieces from hyperlinked Web pages. Furthermore, authors present and experimentally evaluate heuristic algorithms to generate the top composed pages.

It is necessary to note [14] which proposes an effective genetic algorithm that monitors the success of internet database management system by combining functionality, quality, and complexity of a query optimizer for finding good solutions to the problem.

Our research and development are in line with the second trend. Despite the similarity of the question area in the above mentioned studies, our approach to the analysis of the genetic algorithm behavior is different in some features. In particular, we use:

- proprietary specific-reasoned definition of fitness function;

• proprietary interpretation of genetic crossing and mutation operations;

• methods of processing the expert assessments of target parameters of algorithm efficiency (such as DCG).

It is also worth noting that there is a similarity of approaches which consists in encoding of chromosomes as sets of key words and concepts with parameters of their contextual relevance.

3. Genetic Algorithm for Generating Search Queries and Ranking Search Results

3.1. **General Concepts.** The project Distributed Intelligent System for Information Support of Innovation in Science and Education [15], [16], [17] proposes the technology of generating search queries, filtering and ranking search results. The main idea is to organize with a special genetic algorithm an evolutionary process generating a stable and effective system query population for getting highly relevant results. In the course of the process coded queries are sequentially exposed to genetic changes and made in a retrieval system. Then the semantic relevance of intermediate search results is evaluated, target function values are computed, and the most appropriate queries are selected.

The search pattern of document K is a set of key words from text documents being reference ones for the subject search area.

Each search query is coded with vector $\overline{q} = (c_1, c_2, ..., c_n, ..., c_m)$, where $c_n = \{k_n, w_n, S_n\}$, $k_n \in K$ is a term, w_n is a term weight, S_n is a set of synonyms for term k_n. The result of a search query is a set of documents R, $|R| = D$. The set R is grouped after performing $\overline{q}$ in a search engine (Bing, Google, SQL database, XML-data, etc.).

The initial population from N search queries is presented as a set of Q_0, where $|Q_0| = N$, $N < |K|/2$, $\overline{q} \in Q_0$. The crossover(one-point or two-point) is carried out by exchanging the terms between components of vectors $\overline{q}$, genotype outbreeding being used for query reproduction. The most adequate mutation operationis the probabilistic change of query term k_n chosen randomly by synonym $k'_n \in S_n$. To generate a new query population an elite selection is used. Generally, the condition of terminating the algorithm is considered to be population stability.

Table 1 shows the correspondence of some basic concepts of information retrieval and genetic algorithms.

3.2. **Fitness Function.** Value of fitness function or applicability function $\overline{W}$ determines the quality of queries (applicability of population

Table 1. Basic concepts of information retrieval and genetic algorithms

Symbol	Information Retrieval Term	Genetic Algorithm Term
q	Query	Chromosome
$k_n \in K$	Query term	Gene
K	Document search pattern	Gene pool
Q	Set of queries	Genotype
$r_i \in R$	Search query result	Phene
R	All query results	Phenotype
	Exchange of terms in queries	Crossover
	Replacing a term by a synonym	Mutation

individuals); genetic algorithm searches for maximum $\overline{W}$:

$$\overline{W} = \frac{1}{N}\sum_{j=1}^{N} \overline{w}_j \to max \tag{1}$$

where
$\overline{W}$ – fitness function of the population;
N - number of queries in population;
$\overline{w}_j$ - fitness function of the j-th population query:

$$\overline{w}_j = \frac{1}{R}\sum_{i=1}^{R} w_i(g,p,s) \tag{2}$$

where
R – number of search results under review in every query;
w_i -fitness function for the i-th result of the j-th query (result r_i):

$$w_i = w_g * g + w_p * p + w_s * s \tag{3}$$

Value g takes account of rank for r_i set by a search system. It is determined from the following:

$$g = 1 - \frac{g(r_i,R) - g_{min}}{g_{max} - g_{min}} \tag{4}$$

$$g(r_i,R) = \sum_{i=1}^{N} pos(r_i)_j^R \tag{5}$$

where

$pos(r_i)_j^R$ – position number r_i in a ranked list of results of the j-th population query;

g_{max}, g_{min} – maximum and minimum values $g(r_i, R)$ of the all population query results.

Value p takes account of genericity r_i that is frequency of occurrence r_i in the list of results of other queries. It is determined from the following:

$$p = \frac{p(r_i, R) - p_{min}}{p_{max} - p_{min}} \tag{6}$$

$$p(r_i, R) = \sum_{i=1}^{N} count(r_i)_j^R \tag{7}$$

where

$count(r_i)_j^R = 1$, if r_i is in the list of results of the j-th query, otherwise $count(r_i)_j^R = 0$;

p_{max}, p_{min} – maximum and minimum values $p(r_i, R)$ of all the population query results.

Value s determines semantic similarity r_i and search pattern K. We use a cosine semantic similarity measure of document vectors as it is common in a vector space model [18]. Then:

$$s(r_i, K) = \frac{\overline{v}(r_i)\overline{v}(K)}{\parallel \overline{v}(r_i) \parallel \cdot \parallel \overline{v}(K) \parallel} \tag{8}$$

where

$\overline{v}(r_i) = \overline{v}(w_1^r, w_2^r, ..., w_n^r, w_{|T|}^r,)$ – vector of the i-th query result, $|T|$ - number of terms in a query result text after morphological analysis (only nouns and adjectives were used) and lemmatization. Document title and its summary (snippet) are used as a result text;

$w_n^r = tf_n^r * idf_n^r$ – weight of the n-th term from the text of query result;

tf_n^r – freq of term use in a text;

$idf_n^r = \log \frac{R+1}{R^n}$;

R^n – total number of results where texts contain the n-th term of the i-th result;

$\overline{v}(K) = \overline{v}(w_1^K, w_2^K, ..., w_m^K, w_{|K|}^K,)$ – vector of the search document pattern K, $|K|$ – number of terms in K;

$w_m^K = \frac{1}{|K|} * idf_m^K$ – weight of the m-th term of the search document pattern K;

$idf_m^K = \log \frac{R+1}{R^m}$;

R^m – total number of results where texts contain the m-th term of K;

w_g, w_p and w_s – weighting factors for g, p and s correspondingly.

3.3. **Implementation of Genetic Algorithm.** [19] proposes a prototype of a program implementation of the genetic algorithm described. In particular, main algorithm steps and parameters, software components, and the preliminary results of the algorithm study are determined. The prototype is implemented as Genetic Algorithm Framework (GAF).

The experimentation described in the article used GAF consisting of main library, the module of morphological analysis and lemmatization, the module of texts similarity semantic analysis, a search module, a database management module, a metadata management module, a user interface (Fig. 1).

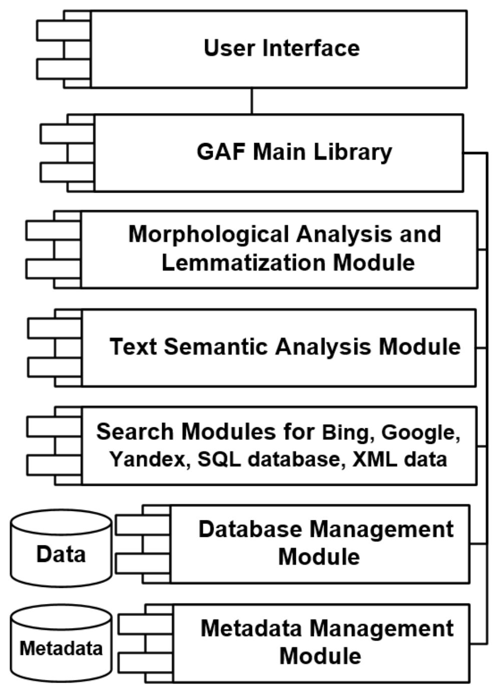

Figure 1. Components of GAF

4. Assessment of Algorithm Convergence

Convergence is the ability of an iterative algorithm to achieve the optimum of goal function or to approach close to it in a finite number of steps. One of the main problems of genetic algorithms intended to solve optimization problems with a large number of local optima is premature convergence. In some cases the problem of premature convergence of algorithms can be solved by the change of parental chromosomes selection strategy in crossing-over; tracking the occurrence of groups of identical chromosomes and their disposal; higher probability of mutation. It is significant that the mathematical justification of why a genetic algorithm finds the optimal solution is not available. In case of simple algorithms, convergence can be proved by the Hollands schema theorem [20].

The results of experiments with the algorithm developed, provided hereinafter, show its behavior with sufficiently high number of iterations. The experimental data allow performing a preliminary assessment of convergence rate under different parameters of fitness function. The following initial set of values of base parameters and a search engine Bing have been used:

- The number of queries in the generated populations $N = 8$.
- The number of keywords in each generated query $M = 6$.
- The number of search results returned by either the query $R_q = 20$, or the query population $R_Q = 20$ or all the populations cumulatively $R = 20$.
- The factor of document arrangement on one server $s_1 = 0.75$.
- Weight factors for arguments g, p and s respectively in ranking a search result $w_g = 0.33$, $w_p = 0.33$, $w_s = 0.34$. The environmental factor a was not taken into account in the experiments.
- The probability of query mutation $p_m = 0.1$.
- The number of the algorithm iteration (or the number of populations generated) $N_Q = 20$.

The initial set K is generated by terms of subject domain related to the control over the technological processes evolution at industrial enterprises; $|K| = 50$.

Fig. 2 shows plots of fitness function value $\overline{W}$ versus a population number with different values of weights w_g, w_p and w_s used in the calculation of $\overline{W}$. In the first case ($w_s = 0.8$) there appears the starting point of algorithm relative stabilization and the achievement of (population 6). In the second case ($w_s = 1.0$) stabilization of results and $\max \overline{W}$ are observed later (population 12). In the third case ($w_s = 0.34$) two local maxima $\overline{W}$ (populations 20 and 8) can be seen, however, it is to early to speak about global maxima $\overline{W}$. Stabilization and $\max \overline{W}$ can be achieved provided the number of algorithm passes increases up to $N_Q = 100$ (Fig. 3). Such dissimilarity from the first two cases can be explained by stronger influence of arguments g and p, with their values being subject to specific algorithms of search engine ranking.

It is notable that there is an almost constant value of the standard deviation values of fitness function in population queries $Var(\overline{w})$ (Fig. 4).

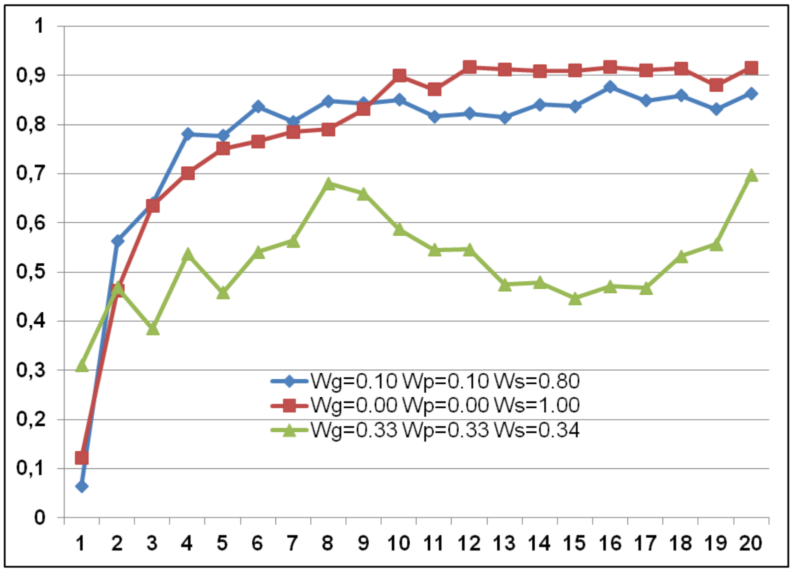

FIGURE 2. Fitness function value $\overline{W}$ versus a population number

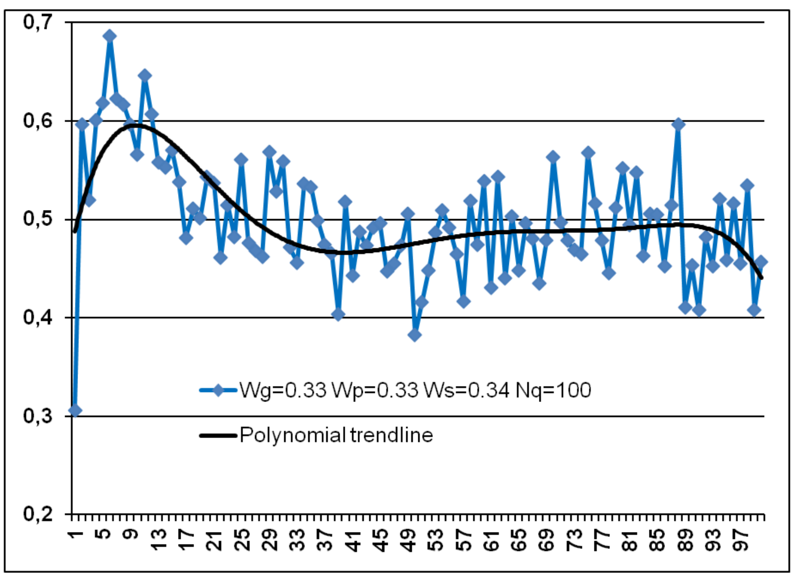

FIGURE 3. Fitness function value $\overline{W}$ versus a population number

This indicates a low probability of significant errors in determining the relevance of documents when using the algorithm.

In general, these experiments clearly demonstrate the convergence of the developed algorithm under different values of its parameters. Though statistical significance of results can be a little bit questionable, we consider it sufficient for positive conclusions.

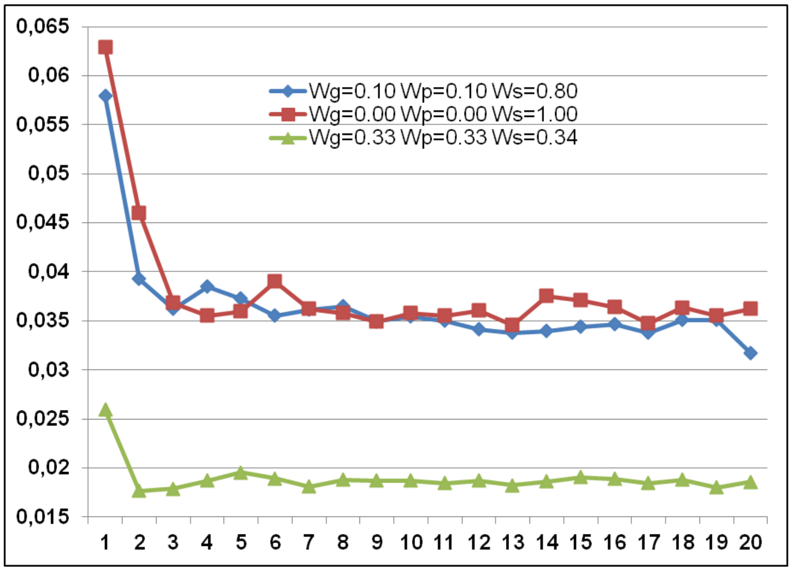

FIGURE 4. Standard deviation values of fitness function in population queries $Var(\overline{w})$ versus a population number

5. Analysis of Search Results Relevance

The goal of our research is to demonstrate the ability of the developed genetic algorithm to provide the high precision of search. Hereinafter we outline the main point of experiments and their results, refine a concept and integrate the data on final statement and conclusion. The detailed picture of experiments and some preliminary results are given in [21].

5.1. **Brief Description of Experiments.** Twelve expert researches of different intellectual fields were invited to take part in the study. Each expert provided a text material (the authors abstract of thesis, a monograph, one or several papers) adequately corresponding to the experts field of knowledge (Table 2).

Besides, all the queries were developed in two ways: with fixed terms and with term lemmas suitable for word forms. In the first case the result identifiers have prefix Q, for example, Bing.Q.

The results of all queries on every subject, both in GAF/Bing and Bing, were transferred to experts in order to estimate their relevance. It is significant that lists of document addresses in the Internet were arranged by the algorithm unknown to experts in order to eliminate influence of searching techniques on the relevance assessment.

Relevance is proposed to be understood as ratio of the information volume useful for subject development to the total amount of information

Table 2. Query types

Query Type ID	**Query Type**	**Query Description**	**Parameters of Executing Queries**	**Search Engine**
Bing	1	A set of key words from title of the experts material	$M = 5...11$; $R = 20$	Bing
GAF/ Bing	2	A set of key words from text of the experts material; $\lvert K\rvert = 50$	$N = 6$; $M = 8$; $R_q = R_Q = R = 20$; $s_1 = 0.75$; $w_g = 0.33$; $w_p = 0.33$; $w_s = 0.34$; $p_m = 0.1$; $N_Q = 10$	GAF environment with the Bing search module (further GAF/Bing environment)

in the material. To assess the document relevance it was proposed to use the following scale: "totally relevant", "partial relevant", "lack of relevance", "not relevant". Moreover relevance was assessed from the viewpoints of highly skilled specialists (experts) and entrants.

5.2. **Discussion.** [16] and [21] make a preliminary conclusion that GAF/Bing search precision is higher than Bing precision and the average precision of the documents retrieved is practically the same. Besides, the search hit ratio of the retrieved document addresses is 2% which means that GAF/Bing method makes it possible to find new original documents. The following presents new advanced experimental data on the comparative evaluation of the retrieved document relevance.

Consider the average relevance of the documents retrieved in the following fields: agricultural technologies, astrophysics, construction materials, databases, e-learning, friction and wear, materials engineering. Fig. 5 shows the total average relevance, Fig. 6 shows it in areas. We can see (Fig. 6) that the number of relevant documents retrieved with different methods for each application area is quite remarkably different. For example, the astrophysics area results showed the best ones with GAF/Bing for searching with fixed terms and the worst ones with the

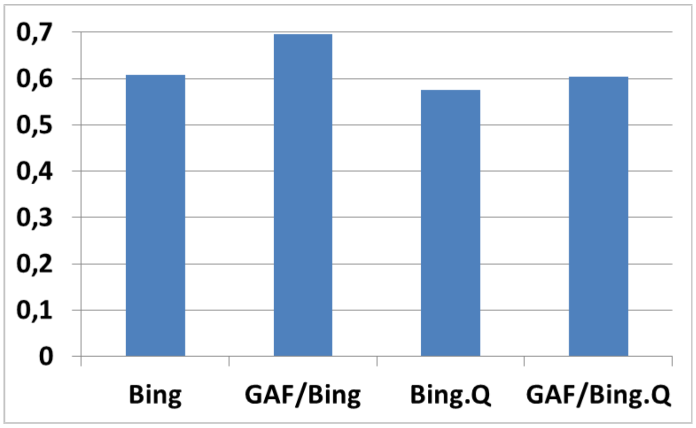

FIGURE 5. Average relevance of search results as a whole in all areas

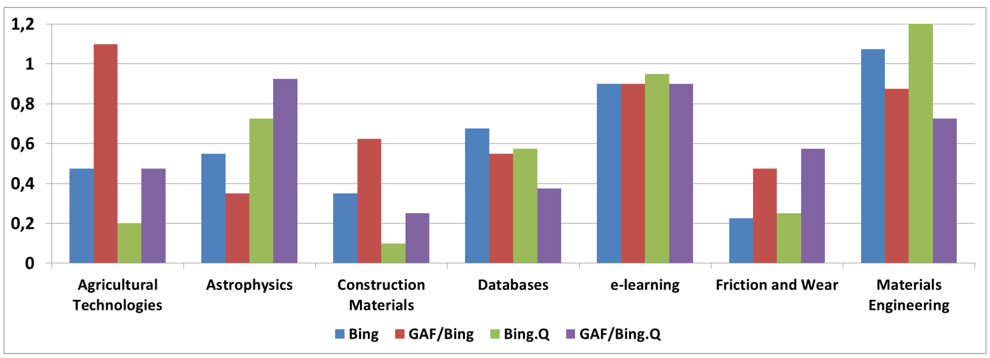

FIGURE 6. Average relevance of search results in areas

same GAF/Bing but with lemmatized terms. The construction materials area, on the contrary, showed the best results with GAF/Bing lemmatized query terms and the worst ones with GAF/Bing fixed query terms. As for the e-learning area the results of all search methods were virtually the same.

Nevertheless, the average values of total search result relevance for all areas are higher with GAF/Bing (approximately 10%).

We interpret these experimental data as follows. It is quite natural that the quality of document relevance expertise cannot be equal as it depends on the experts qualification or, more generally, on the end users skill to select pertinent documents. Another point is the approach taken to forming queries for a search system. At this point key words were not selected specially; we used a typical and trivial way of retrieving key words from the titles of source documents. In addition, we deliberately chose very different application areas (e.g., astrophysics and e-learning) in order to provide the universal search character and minimize the domain intersection of query key words.

Thus, the experts, on the whole, simulated the real life procedure of searching and retrieving relevant documents. Moreover, the procedure results were compared with the results of GAF evolutionary model which is considered to be rather more efficient.

Now consider the search precision, i.e. the ratio of relevant documents in search results. Fig. 7 shows the average search precision, Fig. 8 shows the quantity of relevant documents ranked by relevance in search results.

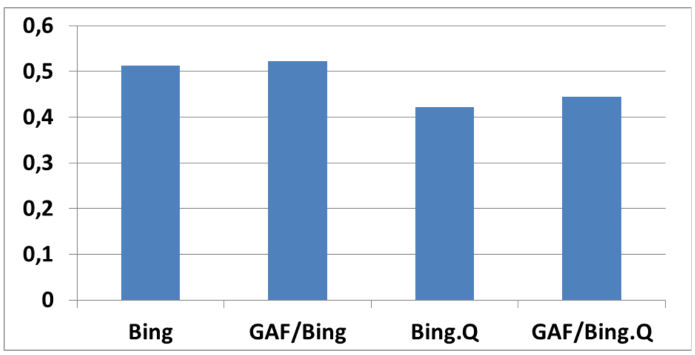

FIGURE 7. Average search precision (the ratio of relevant documents in search results, in general)

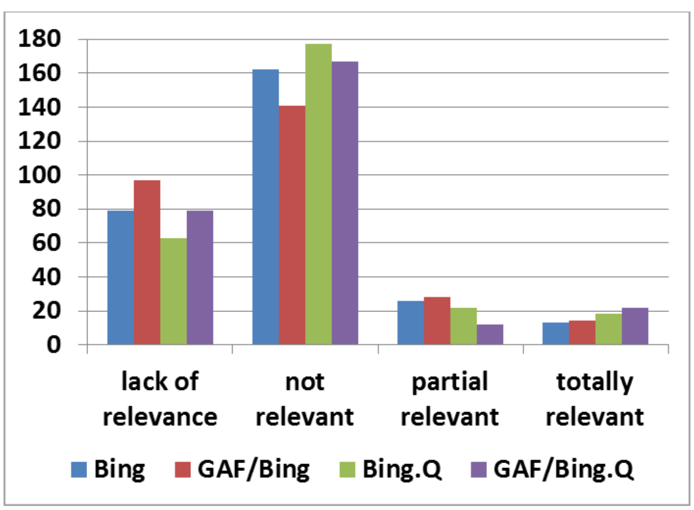

FIGURE 8. Quantity of relevant documents in search results (ranked by relevance)

We can see (Fig. 8) that GAF/Bing and Bing search results give approximately the same quantity of totally relevant and partially relevant documents. However, Bing shows more completely not relevant documents and GAF/Bing gives more documents with lack of relevance.

The average search precision (Fig. 7), in general for all application areas, is higher for GAF/Bing (though not much, approximately 2%) .

How can we explain that? We suppose that additional relevant GAF/Bing results appear through applying the suggested evolutionary method of generating pertinent search query pool. New combinations of key words make it possible to retrieve new documents. The probability of new highly relevant documents is, of course, rather slight; the state-of-the-art algorithms of ranking search systems, one way or another, display such documents at the top of search results. However, the total increase in the quantity of relevant search results, which GAF shows, speaks for the method positive effect.

It should be noted that the relevance expertise is only used as a base for analyzing experimental data and is not used when GAF works. The ranking of GAF search result relevance is done automatically in the course of computing fitness function values $\overline{W}$.

6. Total Findings

Thus, concluding the results of the research we can state the following.

1. The convergence of the developed genetic algorithm is evaluated. The results of the experiment with the designed genetic algorithm show its capability to reach the neighborhood of fitness function optimum with the finite number of algorithm steps. This suggests that the algorithm can find and retrieve really more relevant documents.

2. The search result relevance based on the search precision calculation is evaluated. The results derived show rather higher ratio of relevant documents in search results as compared with commonly known Internet search system. It is to be noted that the search precision is evidently the most significant criterion of search result evaluation.

3. The experimental results described in the paper confirm the preliminary conclusions of [16], [21]. Namely, the fact that the GAF/Bing method allows a user to find new original documents and, in addition, the average relevance of documents found by making use of the proposed GAF/Bing method is almost as good as the average relevance of the documents found by Bing.

7. Conclusion

It is reasonable to assume that the maximum effect of the developed technology is achieved by searching sources of information on the new, for a specialist, subject at early stages of its studying and mastering. Indeed, just that very stage of research needs the analysis and evaluation of the maximal quantity of information resources on the subject under study.

The main direction of further research, in our opinion, is connected with testing the suggested technology with test collections [22]. We are planning to prepare and conduct the research with such commonly known collections as CACM, CISI, INSPEC, and LISA. In addition we are projecting the experiments with data collections TREC such as Web [23] and KBA [24] search.

References

[1] H. Chu, *Information representation and retrieval in the digital age* (American Society for Information Science and Technology by Information Today Inc., Medford, NJ, 2010).

[2] C.D. Manning, P. Raghavan, H. Schutze, *Introduction to information retrieval* (Cambridge University Press, Cambridge, England, 2009).

[3] A. Broder, ACM SIGIR Forum **Vol. 36** (Issue 2), 3-10 (2002).

[4] . Agbele, A. Adesina, D. kong, O. Ayangbekun, Applied Computational Intelligence and Soft Computing **Vol. 2012**, 7 (2012).

[5] S.S. Sathya, Ph. Simon, International Journal of Computer Theory and Engineering **Vol. 1** (Issue 4), 450-455 (2009).

[6] L.C. Chen, C.J. Luh, C. Jou, Information Systems **Vol. 30** (Issue 4), 299316 (2005).

[7] T.P.C. Silva et al, Information Systems, **Vol. 34** (Issue 2), 276289 (2009).

[8] F. Johnson, S. Kumar Advances in Computing, Communication, and Control Communications in Computer and Information Science **Vol. 361**, 82-93 (2013).

[9] M.H. Marghny, A.F. Ali, AIML'05 Conference, CICC, Cairo, Egypt, 82-87 (2005).

[10] L. Tamine, C. Chrisment, M. Boughanem, Information Processing and Management **Vol. 39**, 215231 (2003).

[11] R.L. Cecchini et al, Information Processing and Management, **Vol. 44** 863-878 (2008).

[12] R.L. Cecchini et al, Journal of the American Society for Information Science and Technology, **Vol. 61** (Issue 6), 1258-1274 (2010).

[13] R. Varadarajan, V. Hristidis, IEEE Transactions on Knowledge and Data Engineering, **Vol. 20** (Issue 3), 411-424 (2008).

[14] M. Sinha, S. Chande, Research Journal of Information Technology, **Vol. 2** (Issue 3), 139144 (2010).

[15] V.K. Ivanov, B.V. Palyukh, A.N. Sotnikov, Programmnyye produkty i sistemy, (Issue 4), 197-202 (2013).

[16] V.K. Ivanov, B.V. Palyukh, A.N. Sotnikov, FedCSIS'2014 Annals of Computer Science and Information Systems, **Vol. 3**, 13-20 (2014).

[17] V.K. Ivanov, Innovatsii v nauke, (Issue 25), 8-15 (2013).

[18] G. Salton, A. Wong, C.S. Yang, Communications of the ACM, **Vol. 18** (Issue 11), 613620 (1975).

[19] V.K. Ivanov, P.I. Meskin, Programmnye produkty i sistemy, (Issue 4), 118-126 (2014).

[20] J.H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence* (The MIT Press, 1992).

[21] V.K. Ivanov, B.V. Palyukh, OSTIS-2015 Materialy konferencii, 471-476 (2015).

[22] M. Sanderson, Foundations and Trends in Information Retrieval, **Vol. 4** (Issue 4), 247375 (2010).

[23] K. Collins-Thompson et al, The Twenty-Second Text REtrieval Conference (TREC 2013) Proceedings, **Vol. SP 500-302**, (2014).

[24] J.R. Frank et al, The Twenty-Second Text REtrieval Conference (TREC 2013) Proceedings, **Vol. SP 500-302**, (2014).