Searching the Web in Context: Genetic Algorithms for Exploring Query Space*

Rocío L. Cecchini[†], Carlos M. Lorenzetti[‡],
Ana G. Maguitman[‡] and Nélida Beatriz Brignole[†§]

[†] LIDeCC - Laboratorio de Investigación y Desarrollo en Computación Científica
[‡] LIDIA - Laboratorio de Investigación y Desarrollo en Inteligencia Artificial
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur, Av. Alem 1253, (8000) Bahía Blanca, Argentina
phone: 54-291-4595135 fax: 54-291-4595136

[§]Planta Piloto de Ingeniería Química (UNS-CONICET)
Cno la Carrindanga km 7, (8000) Bahía Blanca, Argentina

e-mail: {cr,cml,agm,nbb}@cs.uns.edu.ar

Abstract. How to enable effective use of on-line information is an increasing problem, as technology enhances the ability to publish information rapidly and large quantities of information are instantly available for retrieval. In text-based web search, users' information needs and candidate text resources are typically characterized by terms. Therefore, the quality of web retrieval is highly dependant on the vocabulary used to generate the search queries. Selecting good query terms is extremely difficult and presents many interesting research challenges. This paper applies Genetic Algorithms to develop a framework for the dynamic identification of "good query terms" to aid web search in the context of a theme. Starting with a population of terms from a thematic context, the techniques presented in this paper incrementally identify good candidate search queries. After a few generations, the population of queries evolves towards a set of queries that allows to retrieve material relevant to the given context. The proposed techniques have the potential to identify good queries even if some of the terms in the queries do not occur in the original context.

1 Introduction

Every day, search continues to grow in popularity and search engines have become key media for many of our daily activities. Meaningful context-based access to the web is crucial to effectively participate in today's information society. Accessing context relevant information through search engines requires the formulation of appropriate queries. Unfortunately, contextualizing web search is

* This research work is supported by Agencia Nacional de Promoción Científica y Tecnológica (PICT 2005 Nro. 32373), Universidad Nacional del Sur and Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).

challenging. In current search engines, there are limits on query length, or if long queries are allowed, they may become too specific, returning very few or no results. This makes it difficult to provide appropriate queries to describe rich contexts. Even if special syntaxes are used to formulate context-based queries, there is no guarantee that the vocabulary used to describe the context will match the vocabulary by which the relevant resources are indexed. The goal of our research work is to design novel techniques to automatically refine search queries and to accumulate resources relevant to a thematic context as a whole.

This paper describes a framework based on Genetic Algorithms (GAs) that addresses the problem of reflecting topical information when formulating search queries. The proposed framework takes a novel incremental approach to evolve high-quality queries for retrieving context-relevant textual resources (such as html pages, pdf files, Word files, etc.). It starts by generating an initial population of queries using terms extracted from a thematic context and incrementally evolves those queries based on their ability to retrieve relevant results when presented to a search engine.

Developing methods to evolve high-quality queries and collect context-relevant resources can have important impacts on today's information society. These methods can help build systems for a range of information services:

- **Task-Based Search.** Task-based search systems exploit user interaction with computer applications to determine the user's current task and contextualize information needs [11,3]. Basic keyword searches could very easily miss task-relevant pages. By evolving high-quality queries, a task-based search system can automatically generate suggestions that are richly contextualized within the user's task.
- **Resource Harvest for Topical Web Portals.** Topical web portals have the purpose of gathering resources on specific subjects. The collected material is used to build specialized search and directory sites. Typically, focused crawlers are in charge of mining the web to harvest topical content and populate the indices of these portals [6,14]. As and alternative to focused crawlers, this process can be supported by formulating topical queries to a search engine and selecting from the answer set those resources that are related to the topic at hand.
- **Deep Web Search.** Most of the web's information can be found in the form of dynamically generated pages and constitute what is known as the deep web (aka hidden web or invisible web) [9,15]. The pages that constitute the deep web do not exist until they are created dynamically as the result of a query presented to search forms available in specific sites (e.g., pubmedcentral.nih.gov, amazon.com). Therefore, the formulation of high-quality queries is of utmost importance at the moment of accessing deep web sources. For that reason, searching the deep web in context is an important area of application for the proposed techniques.
- **Support for Knowledge Management.** Effective knowledge management may require going beyond initial knowledge capture, to support decisions about how to extend previously-captured knowledge [10,12]. The web

provides a rich source of information on potential new material to include in a knowledge model. Thus material can be accessed by means of contextualized queries presented to a conventional search engine, where the context is given by the knowledge model under construction. Using the web as a huge repository of collective memory and starting from an in-progress knowledge model, the techniques discussed here can facilitate the process of capturing knowledge to help extend organizational memories.

In the next section we present an overview of GAs and discuss their suitability for their application to our research problem. Then we present the main contribution of this paper, a GA approach for evolving high-quality queries, followed by an evaluation of the proposal. The paper closes with a summary of our conclusions and a discussion of future work.

2 Background

2.1 Genetic Algorithms

GAs [8] are robust optimization techniques based on the principle of natural selection and survival of the fittest, 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.

To use GAs in optimization problems we need to define candidate solutions by chromosomes consisting of genes and a fitness function to be maximized. A population of candidate solutions (usually of a constant size) is maintained. The goal is to obtain better solutions after some generations. To produce a new generation GAs typically use selection together with the genetic operators of crossover and mutation. Parents are selected to produce offspring, favoring those parents with highest values of the fitness function. Crossover of population members takes place by exchanging subparts of the parent chromosomes (roughly mimicking a mating process), while mutation is the result of a random perturbation of the chromosome (e.g., replacing a gene by another). A simple GA works as follows:

Step 1: Start with a randomly generated population
Step 2: Evaluate the fitness of each individual in the population
Step 3: Select individuals to reproduce based on their fitness
Step 4: Apply crossover with probability Pc
Step 5: Apply mutation with probability Pm
Step 6: Replace the population by the new generation of individuals
Step 7: Go to step 2

Although selection, crossover and mutation can be implemented in many different ways, their fundamental purpose is to explore the search space of candidate solutions, improving the population at each generation by adding better offspring and removing inferior ones.

2.2 GAs for Context-Based Web Search

The practical motivation for our research work is the development of web search tools for collecting resources relevant to a thematic context. Context-based retrieval systems create a model of the user context, infer user needs for additional information resources, and search for relevant documents on the web or other online electronic libraries. Traditionally, such systems find documents relevant to the user information needs by augmenting user queries with other terms selected from the context. A variety of systems pursuing this approach have obtained encouraging results (e.g. [16,3,13]).

There are a number of reasons why GAs are appropriate to deal with the problem of context-based web search:

- **Context-Based Web Search as an Optimization Problem.** Generating high-quality queries for context-based search on the web can be regarded as an optimization problem. The search space of the problem is defined as the set of possible queries that can be presented to a search engine. The objective function to be optimized is based on the effectiveness of a query to retrieve relevant material when presented to a search engine. Depending on the system goals, a measure of query effectiveness can be defined using traditional IR notions such as precision and recall, or other customized performance evaluation metrics.
- **High Dimensional Space.** Query space is a high dimensional space, where each possible term accounts for a new dimension. This kind of problems cannot be effectively solved using analytical methods but are natural for GAs.
- **Suboptimal Solution.** Successful web search requires the formulation of high-quality queries even if the formulated queries are not the optimal ones. GAs do not guarantee the identification of optimal solutions but are usually successful in finding near optimal ones.
- **Multiple Solutions.** Each one of multiple sets of web pages can represent a satisfactory result for a context-based search. Therefore, we may be interested in finding many high-quality queries rather than a single one. GAs can be naturally used for multimodal relevance optimization.
- **Exploration and Exploitation.** Finding good combinations of query terms requires exploring different direction of the thematic-context space. This exploration must be independent of the initial population of queries and it may require going beyond the initial set of terms by incorporating novel terms. Such a search process can be effectively performed by applying the genetic operators of crossover and mutation. In addition, the exploitation of the most promising combination of terms is naturally induced by the selection mechanism.

3 A Genetic Approach for Evolving High-Quality Queries

The goal of this research work is to evolve queries that have the capability of retrieving material similar to the user context when presented to a search in-

terface. In order to accomplish this goal we start with a population of queries composed of terms extracted from the user context and rate each query according to the quality of the search results. As generations pass, queries associated with improved search results will predominate. Furthermore, the mating process continually combines these queries in new ways, generating ever more sophisticated solutions. In particular, the mutation mechanisms can be implemented in such a way that novel terms, i.e., terms that are not in the initial user context, are brought into play.

3.1 Population and Representation of Chromosomes

The search space Q is constituted by all the possible queries that can be formulated to a search engine. Thus the population of chromosomes is a subset of such queries. Consequently, each chromosome is represented as a list of terms, where each term corresponds to a gene that can be manipulated by the genetic operators. The population is initialized with a fixed number of queries randomly generated with terms from the thematic context. The number of terms in each of the initial queries will be random, with a constant upper bound on the query size. While all terms in the initial population of queries come from the initial thematic context, novel terms can be included in the queries after mutation takes place. These novel terms are obtained from a *mutation pool*, which is an ever increasing set of terms that may or may not be part of the initial context.

3.2 Fitness Function

We associate with the search space Q a fitness function $\text{Fitness} : Q \rightarrow [0 \ldots 1]$ which can numerically evaluate individual queries. The fitness function defines the criterion for assessing the quality of a query. Our conception of high-quality query is based on the query's ability to retrieve material similar to the thematic context c when submitted to a search engine. The function we propose to measure fitness is

$$\text{Fitness}(\mathbf{q}) = \max_{d_i \in \mathbf{A_q}} (\sigma(c, d_i))$$

where $\mathbf{A_q}$ is the answer set for query $\mathbf{q}$ (set of documents returned by a search engine when $\mathbf{q}$ is used as a query) and $\sigma : D \times D \rightarrow [0 \ldots 1]$ is the similarity measure for a pair of documents (note that the context c can be regarded as a document in D).

Different similarity measures, such as the standard cosine similarity or Jaccard similarity [1], can be used in the implementation of the fitness function. One pragmatic difficulty is the use of the complete answer set $\mathbf{A_q}$ in our definition of fitness. Looking at the entire set of pages returned by a search engine is too expensive for practical purposes. Therefore, we only look at the top ten results and only the "snippets" returned by the search engine are used for computing similarity. (The snippet is a text excerpt from the page summarizing the context where the search terms occur.)

3.3 Genetic Operators

A new generation in a our GA is determined by a set of operator that select, recombine and mutate queries of the current population.

- **Selection:** A new population is generated by probabilistically selecting the highest-quality queries from the current set of queries. The probability that a query $\mathbf{q}$ will be selected is proportional to its own fitness $f(\mathbf{q})$ and is inversely proportional to the fitness of the other queries in the current population. This method is known as the roulette-wheel selection.
- **Crossover:** Some of the selected queries are carried out into the next generations as they are, while others are recombined to create new queries. The recombination of a pair of parent queries into a pair of offspring queries is carried out by copying selected terms from each parent into the descendants. The crossover operator used in our proposal is known as single-point. It results in new queries in which the first n terms are contributed by one parent and the remaining terms by the second parent, where the crossover point n is chosen at random.
- **Mutation:** Small random changes can be produced to the new population of queries. These changes consist in replacing a randomly selected query term t^q by another term t^p . The term t^p is obtained from a *mutation pool* (described next).

3.4 Mutation Pool

The mutation pool is a set of terms that initially contains terms extracted from the thematic context under analysis. As the system collects relevant content, the mutation pool is updated with new terms from the snippets returned by the search engine. This procedure brings new terms to the scene, allowing a broader exploration of the search space.

3.5 Proposed System Architecture

Figure 1 depicts the proposed system architecture for a contextualized web-search system based on GAs. In the proposed prototype, the system will maintain an internal representation of the thematic context. In addition it will maintain a population of queries which is incrementally refined as the system evolves. The basic mechanisms that enable the system to evolve queries and retrieve context-based results are the following:

- **Context Modeling**. This mechanism generates a model of the thematic context under analysis. For example, for a task-based search service, it can observe how the user interacts with different kinds of computer utilities such as email systems, browsers and text editors, and generates a representation of the user's thematic context.

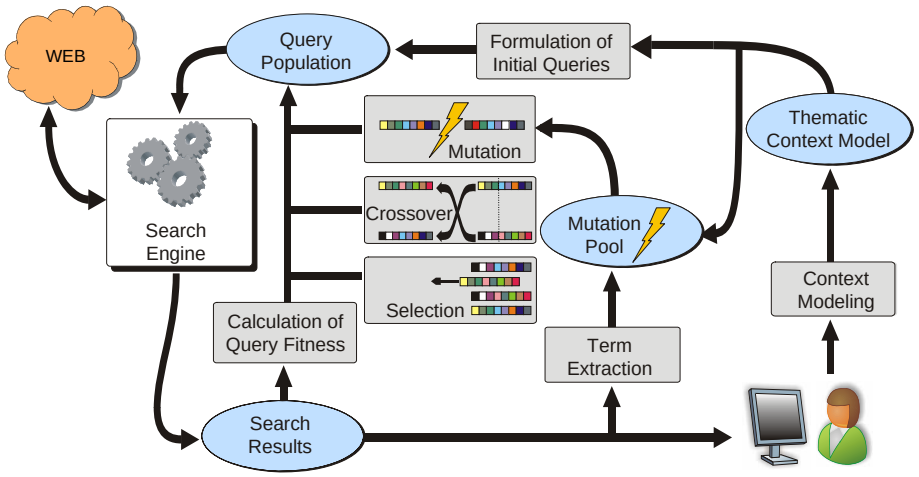


Fig. 1. Architecture for a contextualized web-search system based on GAs.

- **Formulation of Initial Queries**. It selects terms from the thematic context and forms suitable queries, which are submitted to a standard search engine (e.g, Google) or entered into individual search forms (e.g., Amazon or PubMed). Initial queries will be automatically formed using a random selection of terms from the thematic context. The sizes of the initial queries are never more than a predefined constant.
- **Calculation of Query Fitness**. This mechanism estimates the relevance of the results returned by a search engine after submitting a query. Based on the estimated relevance it will associate a fitness value with the query. One way the relevance of a search result can be approximated is by computing the similarity between the collected material and the thematic context, but other approaches can be taken.
- **Term Extraction**. This component uses the content returned by a search engine to extract new terms, which are used to update the mutation pool.
- **Selection, Crossover and Mutation**. These mechanisms, described in section 3.3, are in charge of selecting, recombining and mutating the queries of the current population.

Although the sizes of the initial queries are never more than a predefined constant, the sizes of some queries in subsequent generations can exceed this limit. This is because applying the crossover operator can change the offspring size. Notice that existing search engines use up to a fixed number of terms and ignore subsequent ones (e.g., Googles query size limit is 32 terms). Interestingly, the eventual increase of query size beyond this limit captures, in a rough sense, the phenomenon of recessive inheritance: some terms that are ignored in a generation (because they occur beyond the query size limit) may be taken into account

in subsequent generations when these terms become part of an offspring query after crossover takes place.

4 Evaluation

4.1 Evaluation Criteria

To evaluate the performance of context-based retrieval based on GAs we first had to establish evaluation criteria suitable for this task. We adopted evaluation criteria based on the quality of the best queries at each generation, and the performance improvement is measured as the increase in the quality value as the generations pass.

In order to propose a measure of query quality we first give a precise definition of similarity between a thematic context and a retrieved result. Assume c is a thematic context and $\mathbf{q}$ a query associated with c. Let $\mathbf{A_q} = \{a_1, \ldots, a_n\}$ be the set of retrieved resources (answer set) for $\mathbf{q}$. A measure of similarity between c and a_i can be computed using the *cosine similarity* defined as:

$$\sigma(c, a_i) = \frac{\overrightarrow{c} \cdot \overrightarrow{a}_i}{\|\overrightarrow{c}\| \cdot \|\overrightarrow{a}_i\|}$$

where $\overrightarrow{c}$ is the vector representation of the thematic context based on the terms in c, and $\overrightarrow{a}_i$ is the vector representation of a_i based on the terms occurring in the corresponding snippet returned by a search engine.

We use σ to define *query quality based on maximum similarity* as follows:

$$\text{Quality\_Max}(\mathbf{q}) = \max_{a_i \in \mathbf{A_q}} (\sigma(c, a_i)).$$

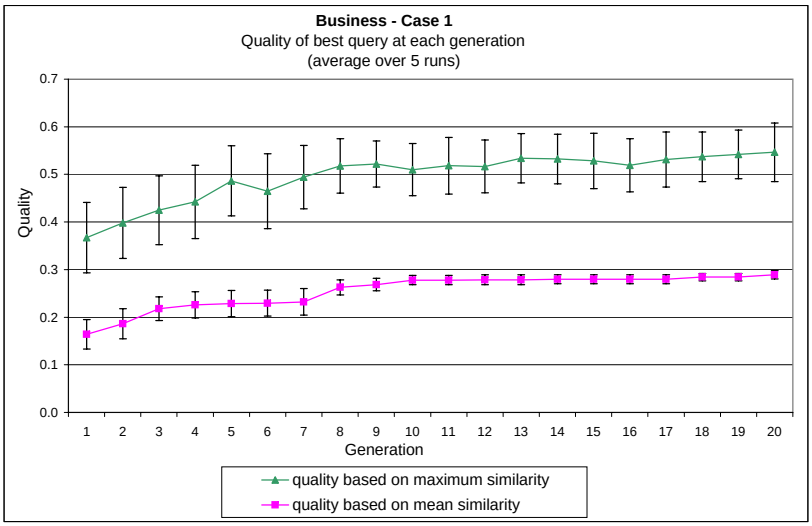
Analogously, we define *query quality based on mean similarity* as:

$$\text{Quality\_Mean}(\mathbf{q}) = \frac{\sum_{a_i \in \mathbf{A_q}} (\sigma(c, a_i))}{|A_q|}$$

Notice that the function Quality_Max is defined exactly as the fitness function presented in section 3.2. On the other hand, Quality_Mean is computed as the average similarity over all pairs (c, a_i). Depending on the task at hand, one notion of query quality may be preferred over the other. For instance, Quality_Max is more appropriate if the goal is to retrieve a unique highly relevant result. Alternatively, Quality_Mean combines the relevance of a set of results and therefore is more appropriate if several results are expected to be useful.

4.2 The Performance Evaluation

A performance test based on our criterion functions requires access to a thematic context c. We generated six thematic contexts to conduct six tests by selecting three topics from the DMOZ directory (dmoz.org) and two webpages from each



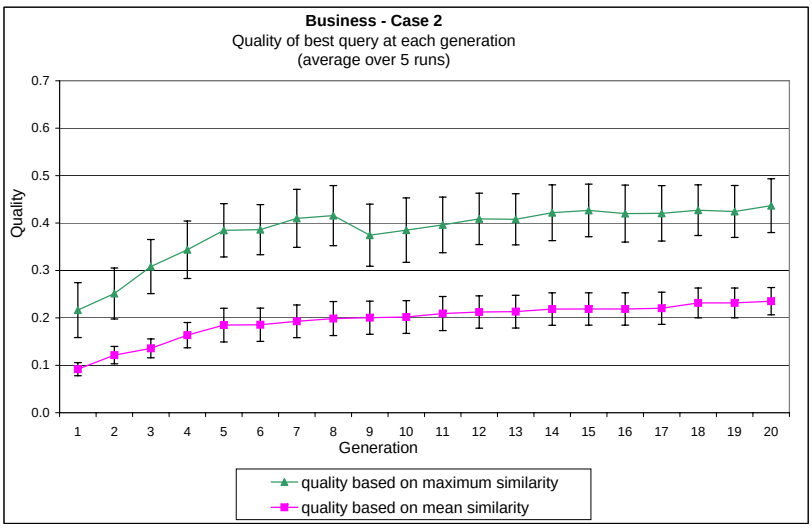
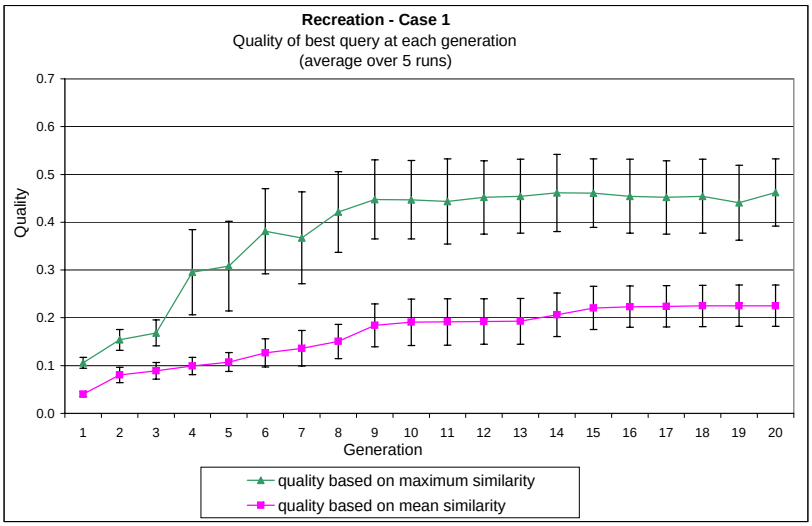


Fig. 2. Two tests showing the average query quality over five independent runs for the topic *Business*.

topic. The topics selected for our tests are *Business*, *Recreation* and *Society*. Each of our six tests consisted in running the GA five times. Each run consisted in 20 generations, with a population of 60 queries, a crossover probability of 0.7 and a mutation probability of 0.03. The population of queries was randomly initialized using the thematic context. The size of each query was a random number between 1 and 32.



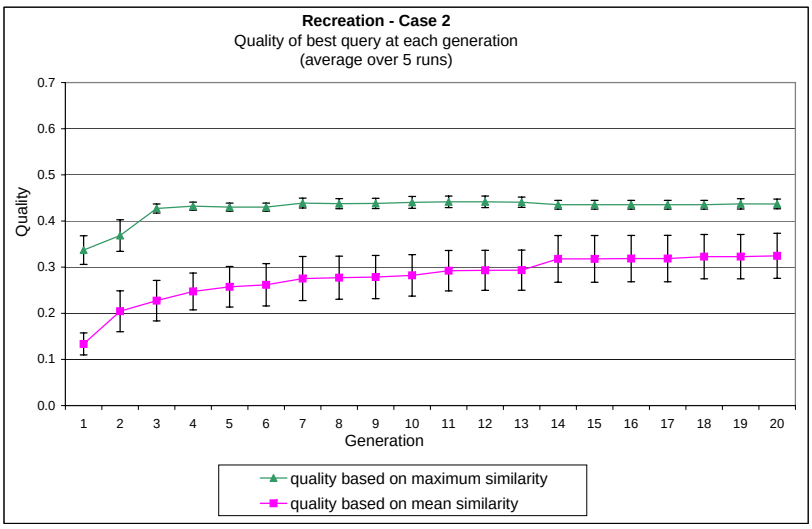
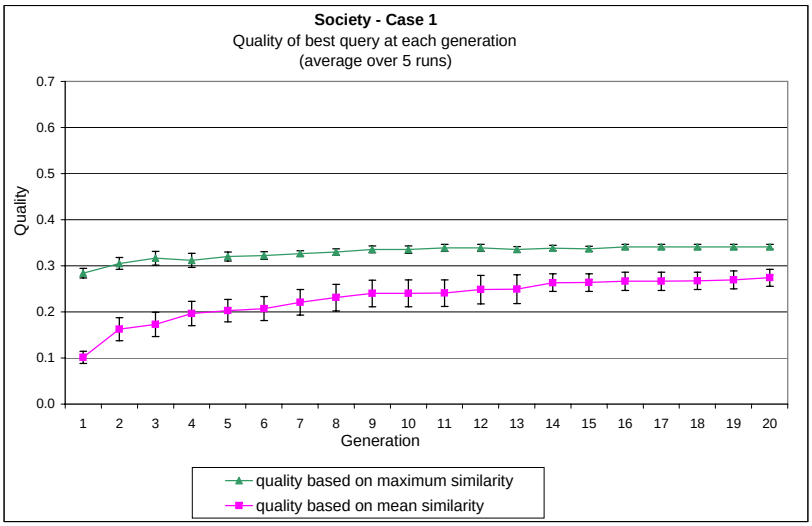


Fig. 3. Two tests showing the average query quality over five independent runs for the topic *Recreation*.

Figures 2, 3 and 4 show the performance of the GA for our three topics. For each generation, we plotted the average quality of the best query (using both Quality_Max and Quality_Mean) and error bars (at 95% C.I.) resulting from the five runs. In all the tests, the comparison of the query quality obtained through a small number of generations shows that the GA results in statistically significant improvements over the initial generations. (Notice that in all our tests the error



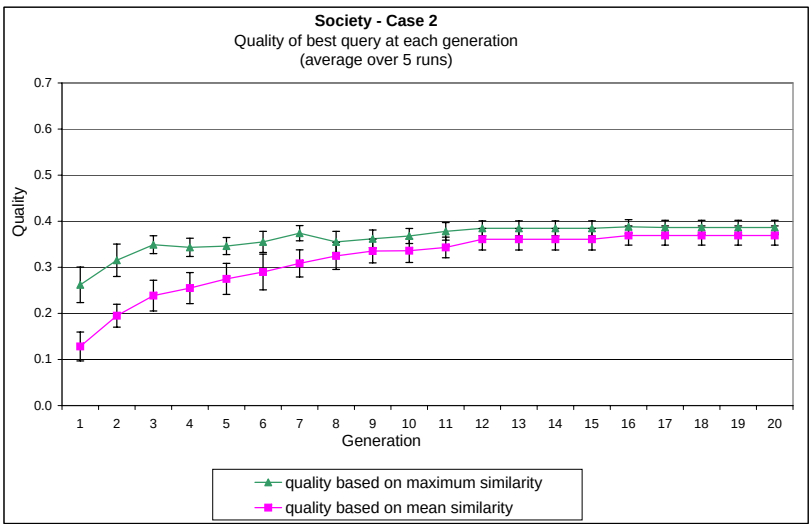


Fig. 4. Two tests showing the average query quality over five independent runs for the topic *Society*.

bar corresponding to the first generation does not overlap with the error bar at some later generation.) In other words, the GA is able to evolve queries with quality considerably superior to that of the queries generated directly from the thematic context.

5 Conclusions and Future Work

There have been some previous proposals to apply GA techniques to deal with problems in the area of information retrieval. Among the existing proposals we can mention the application of GA techniques to derive better document descriptions [7] and for term weight reinforcement in query optimization [18,2]. These proposals differ from ours in attempting to tune the weights of individual terms rather than evolving queries. In addition, while our approach is fully automatic, others require relevance feedback from the users.

The techniques presented in this paper are applicable to any domain for which it is possible to generate term-based characterizations of a context. However, query adaptation involves submitting each new query to a search engine to calculate its fitness, which is a time consuming process. Therefore, we expect the proposed techniques to have potential applicability to exploiting thematic context for non-real time systems, where slow response times are acceptable.

Our initial evaluations show the effectiveness of GA techniques for query generation and refinement. More work, however, needs to be done in this area to make the results richer. We plan to test different settings for the GA parameters (population size, crossover probability and mutation probability). We have used roulette-wheel selection in the implementation of our methods. However, it is known that other selection methods such as tournament selection are better than roulette-wheel at maintaining diversity. Therefore, we plan to study the impact that other selection methods have on the overall performance of our techniques.

Many search engines allow the formulation of queries with special syntaxes that help results get more specific [5]. Consequently, an interesting followup study concerns applying genetic programming to evolve queries that take advantage of these special syntaxes. In such methods not only terms will be important at the moment of formulating queries, but boolean operators and other special commands will be considered as well.

There is also much to investigate regarding the fitness function. Choosing a good fitness function is one of the most important aspects in the development of GAs. We based our definition of fitness on the notion of similarity. This was done to keep in line with classical information retrieval systems that typically attempt to match requests with the most similar documents. However, a few information retrieval approaches take a different position [4,17,12] and postulate that in some circumstances conventional notions of similarity may not be the best criteria for retrieval. In certain scenarios, attaining novelty and diversity may be as important, or even more important, than attaining similarity. Therefore, alternative fitness functions can be defined depending on the task at hand.

References

1. Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

2. Mohand Boughanem, Claude Chrisment, and Lynda Tamine. On Using Genetic Algorithms for Multimodal Relevance Optimization in Information Retrieval. *J. Am. Soc. Inf. Sci. Technol.*, 53(11):934–942, 2002.
3. Jay Budzik, Kristian J. Hammond, and Larry Birnbaum. Information Access in Context. *Knowledge Based Systems*, 14(1–2):37–53, 2001.
4. Jay Budzik, Kristian J. Hammond, Larry Birnbaum, and Marko Krema. Beyond similarity. In *Proceedings of the 2000 Workshop on Artificial Intelligence and Web Search*. AAAI Press, 2000.
5. Tara Calishain and Rael Dornfest. *Google Hacks. 100 Industrial-Strengths Tips and Tools*. O'Reilly, 2003.
6. Soumen Chakrabarti, Martin van den Berg, and Byron Dom. Focused Crawling: a New Approach to Topic-Specific Web Resource Discovery. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1623–1640, 1999. 1999a.
7. M. Gordon. Probabilistic and Genetic Algorithms in Document Retrieval. *Commun. ACM*, 31(10):1208–1218, 1988.
8. John H. Holland. *Adaptation in Natural and Artificial Systems*. Ann Arbor: The University of Michigan Press, 1975.
9. Henry Kautz, Bart Selman, and Mehul Shah. The hidden Web. *AI Magazine*, 18(2):27–36, 1997.
10. David Leake, Ana Maguitman, Thomas Reichherzer, Alberto Cañas, Marco Carvalho, Marco Arguedas, Sofia Brenes, and Tom Eskridge. Aiding Knowledge Capture by Searching for Extensions of Knowledge Models. In *Proceedings of KCAP-2003*. ACM Press, 2003.
11. David B. Leake, Travis Bauer, Ana Maguitman, and David C. Wilson. Capture, Storage and Reuse of Lessons about Information Resources: Supporting Task-Based Information Search. In *Proceedings of the AAAI-00 Workshop on Intelligent Lessons Learned Systems. Austin, Texas*, pages 33–37. AAAI Press, 2000.
12. Ana Maguitman, David Leake, and Thomas Reichherzer. Suggesting Novel but Related Topics: Towards Context-Based Support for Knowledge Model Extension. In *Proceedings of IUI-2005*, pages 207–214, New York, NY, USA, 2005. ACM Press.
13. Ana Maguitman, David Leake, Thomas Reichherzer, and Filippo Menczer. Dynamic Extraction of Topic Descriptors and Discriminators: Towards Automatic Context-Based Topic Search. In *Proceedings of CIKM-2004*, Washington, DC, November 2004. ACM Press.
14. Filippo Menczer, Gautam Pant, and Padmini Srinivasan. Topical Web Crawlers: Evaluating Adaptive Algorithms. *ACM Trans. Inter. Tech.*, 4(4):378–419, 2004.
15. Alexandros Ntoulas, Petros Zerfos, and Junghoo Cho. Downloading Textual Hidden Web Content through Keyword Queries. In *Proceedings of JCDL-2005*, pages 100–109, New York, NY, USA, 2005. ACM Press.
16. Bradley Rhodes and Thad Starner. The Remembrance Agent: A Continuously Running Automated Information Retrieval System. In *Proceedings of PAAM-1996*, pages 487–495, London, UK, April 1996.
17. Barry Smyth and Paul McClave. Similarity vs. Diversity. In *Proceedings of the 4th International Conference on Case-Based Reasoning. Vancouver, Canada*, 2001.
18. Jing-Jye Yang and Robert Korfhage. Query Optimization in Information Retrieval using Genetic Algorithms. In *Proceedings of the 5th International Conference on Genetic Algorithms*, pages 603–613, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.