An Entropy-Based Approach for Preserving Diversity in Evolutionary Topical Search

Cecilia Baggio, Rocío L. Cecchini, Carlos M. Lorenzetti, and Ana G. Maguitman

Institute for Computer Science and Engineering (ICIC)
Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
Universidad Nacional del Sur
Av. Alem 1253 – B8000CPB Bahia Blanca – Argentina
{cb,rlc,cml,agm}@cs.uns.edu.ar

Abstract. Topic-based information retrieval is the process of matching a topic of interest against the resources that are indexed. An approach for retrieving topic-relevant resources is to generate queries that are able to reflect the topic of interest. Multi-objective Evolutionary Algorithms have demonstrated great potential to deal with the problem of topical query generation. In an evolutionary approach to topic-based information retrieval the topic of interest is used to generate an initial population of queries, which is evolved towards successively better candidate queries. A common problem with such an approach is poor recall due to loss of genetic diversity. This work proposes a novel strategy inspired on the information theoretic notion of entropy to favor population diversity with the aim of attaining good global recall. Preliminary experiments conducted on a large dataset of labeled documents show the effectiveness of the proposed strategy.

Keywords: Query reformulation, Information retrieval, Topic-based search, Diversity Preservation, Multi-objective Evolutionary Algorithms

1 Introduction

Topical search is the process of seeking material related to a topic of interest [7]. This process can be automated by formulating and reformulating topical queries in such a way that new queries become increasingly more effective. If documents relevant to the topic of interest are provided as part of a training set, the problem of learning to generate topical queries can be formulated as an optimization problem, where the objective is to maximize effectiveness measures such as precision and recall.

This optimization problem has a number of salient features: (1) the possible solutions are queries that can be formulated by combining all possible words, resulting in a high-dimensional search space, (2) the solutions cannot be constructed efficiently from optimal solutions of its subproblems, (3) we may be interested in finding several quasi-optimal solutions instead of a single optimal one, and (4) several possibly conflicting objectives are involved and therefore the problem can be stated as a Multi-Objective Optimization Problem (MOOP).

In the light of these features, this optimization problem can be naturally approached by the use of Evolutionary Algorithms (EAs) [14], more specifically by Multi-Objective Evolutionary Algorithms (MOEAs) [10, 8]. In previous research work [5, 6] MOEAs have been successfully applied to address this optimization problem by using precision and recall of the individual queries as the objectives to be maximized. However,

for many topical-search applications, queries should be evaluated collectively rather than independently from each other. Therefore, besides attempting to achieve high performance for the individual queries, it is important to attain good recall at the global population level. This situation leads to a new optimization scenario, where diversity preservation becomes an important objective.

This work proposes a novel MOEA strategy that attempts to simultaneously attain topical relevance and diversity. The main feature of the proposed strategy is the use of a novel fitness function, where query performance not only depends on the results retrieved by the individual queries but also depends on the results returned by other queries in the same population. The new metric is a reformulation of the classical recall metric and we refer to it as *entropic recall*. The goal of the new metric is to favor diversity by penalizing queries that retrieve results that are also retrieved by other queries in the same population. This new measure is used in combination with the precision metric to simultaneously favor accuracy and global recall.

The set of possible solutions obtained by the proposed strategy are evaluated using classical performance metrics as well as ad-hoc ones specifically designed to asses the diversity and coverage of relevant results retrieved by the entire population of queries. Therefore, besides assessing the average precision of the generated queries, we evaluate the *global recall* and *diversity* of the entire population of queries. It is important to notice that by evaluating population diversity we do not analyze the diversity among the queries themselves (genotypic level), instead we study diversity among the relevant results retrieved by each query (phenotypic level).

2 A Multi-Objective Evolutionary Algorithm Framework

In this work, we adopt the MOEA framework presented in [6] with the key distinction of using a novel fitness function to favor diversity. The implemented framework applies the Non-dominated Sorting Genetic Algorithm (NSGA-II) [11] selection method to evolve queries guided by the objectives of attaining high precision at the individual query level and high global recall at the population level.

The proposed strategy starts with a population of queries composed of a list of terms extracted from an initial description of the given topic. Each query is interpreted as a disjunctive boolean query. The number of terms in each of the initial queries will be random, with a constant upper bound on the query size. The *decision variable space* Q contains all possible queries that can be formulated to a search interface. Each population of queries is an element of $\mathbb{N}^Q$ (note that repetitions are possible). The documents retrieved by each query are ranked by their similarity to the query. The mating process continually combines these queries in new ways, generating new solutions. As generations pass, the most effective queries will predominate. 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.

Let P be a population of queries, let D_t be the set containing all the documents associated with topic t, including those in its subtopics, let A_q be the answer set returned by a search engine when q is used as a query, and let A_q^{10} be the set of top-10 ranked documents in A_q (note that the size of A_q^{10} might be less than 10 if the search engine

returns less than 10 results for the query q). We adopt *precision at rank 10* as the fitness function used to measure retrieval accuracy. To define *precision at rank 10* we associate with the search space Q and topics T a function that can numerically evaluate an individual query q for a given topic t as follows:

$$\textit{Precision@10}(q,t) = \frac{|A_q^{10} \cap D_t|}{|A_q^{10}|}.$$

The *recall* measure is used as a fitness function to measure coverage at the query level and is defined as follows:

$$\textit{Recall}(q,t) = \frac{|A_q \cap D_t|}{|D_t|}.$$

In order to penalize those queries that do not help to achieve diversity we define a novel fitness function, namely *entropic recall*. The definition of this fitness function has been inspired on the information theoretic notion of *entropy*. The rationale behind the *entropic recall* measure is that a query is penalized if it tends to retrieve the same resources as other queries in the same population. This measure will favor highly discriminative queries, i.e., those queries that can retrieve those resources that other queries are unable to retrieve. Therefore, the output of the *entropic recall* fitness function for a given individual q_i not only depends on q_i but also depends on the resources retrieved by other individuals in the same population. We define the *entropic recall* fitness function as follows:

$$\textit{Entropic-Recall}(q,t,P) = \frac{\sum_{d_i \in A_q \cap D_t} (IQF(d_i,P)/\log(|P|+1))}{|D_t|},$$

where $IQF(d_i,P) = \log((|P|+1)/n_i)$ represents the *inverse query frequency* of document d_i over the collection of documents returned by all the queries in population P, where n_i is the number of queries retrieving d_i. Numerically, the IQF of a document in a population of queries is a log estimate of the inverse probability that a random query q from a population of queries of size $|P|+1$ would return document d_i (1 is added to the population size to ensure $IQF(d_i,P) > 0$). The *entropic recall* measure assesses the *uniqueness factor* of a query q by computing the sum over the IQF of all the relevant documents retrieved by q.

A new generation in our EAs is determined by a set of operators that select, recombine and mutate queries of the current population. The crowded tournament selection operator is used, based on the dominance between two individuals. If the two individuals do not interdominate, the selection is made based on crowding distance. The recombination of a pair of parent queries into a pair of offspring queries is carried out by deleting the crossover fragment of the first parent and then inserting the crossover fragment of the second parent. The second offspring is produced in a symmetric manner. The crossover operator used in our proposal is known as single-point. Our analysis will consider the application of crossover at a typical probability rate (P=0.7). Small changes can be produced to the new population of queries. These changes consist in replacing a randomly selected query term t^i by another term t^j. The term t^j is obtained

from a *mutation pool*. The mutation pool is a set that initially contains terms extracted from the topic description. As the system collects relevant content, the mutation pool is updated with new terms from the relevant documents that have been recovered. This procedure brings new terms to the scene, allowing a broader exploration of the search space. Mutation has been applied at a typical probability rate (P=0.03).

3 Evaluating the Diversity-Preserving Strategies

3.1 Data Collection and Experimental Setup

To run our evaluations we collected the URLs associated with 448 topics from the Open Directory Project (ODP – http://dmoz.org). A number of constraints were imposed on this selection with the purpose of ensuring the quality of our corpus (see [6] for a detailed description of the dataset). The total number of collected pages was more than 350.000. The Lucene framework (http://lucene.apache.org/) was used to index these pages and to run our experiments. We used the stopword list provided by Lucene and Porter stemming was performed on all terms. We divided each topic in such a way that two thirds of its pages were used to create a training index and the remaining one third of the corpus was used for testing. The preliminary evaluations reported in this paper were carried out using the topics BODY PAINTING (168 pages) and AQUACULTURE (362 pages).

The training set was used to evolve queries using 20 independent runs. The number of generations for each run was set to 150 while the population consisted of 100 queries. The test set was then used to determine if the evolved queries for a particular topic were effective on a new corpus. Note that the training and test sets contain the same topics (and therefore the same topic descriptions) but different documents (pages).

With the purpose of evaluating the effect of the *entropic recall* metric we tested two strategies based on the following combinations of fitness functions:

Co1: *Precision@10* and *Recall*.
Co2: *Precision@10* and *Entropic-Recall*.

Combination **Co1** involves classical measures of performance, and therefore was taken as a baseline for comparison purposes.

The performance analysis of the two strategies was completed using on the following metrics:

– $\overline{Precision@10}$. The *average precision at rank 10* is the arithmetic mean of the *precision at rank 10* evaluated across all queries in the population P. Given a population P and a topic t this metric is computed as follows:

$$\overline{Precision@10}(P,t) = \frac{\sum_{q \in P} Precision@10(q,t)}{|P|}.$$

– *Global-Recall*. The *global recall* is the *recall* at the global (population) level. Let the set $A(P)$ be the union of the results returned by a population of queries P:

$$A(P) = \bigcup_{q_i \in P} A_{q_i},$$

where A_{q_i} is defined in the usual way as the set of results returned by q_i. Given a query population $P \in \mathbb{N}^Q$ and a topic t, we compute *Global-Recall* as follows:

$$\textit{Global-Recall}(P,t) = \frac{|A(P) \cap D_t|}{|D_t|}.$$

Global recall measures the success of the entire population of queries in achieving high recall.

– *Jaccard-Similarity-Index*. The *average Jaccard similarity index* is computed by averaging the Jaccard similarity between all pairs of answer sets restricted to relevant documents. Let $A^\star_{q_i} = A_{q_i} \cap D_t$. Then, we compute $\overline{\textit{Jaccard-Similarity-Index}}(P)$ as follows:

$$\overline{\textit{Jaccard-Similarity-Index}}(P) = \frac{\sum_{q_i,q_j \in P, i \neq j} \textit{Jaccard-Similarity-Index}(A^\star_{q_i}, A^\star_{q_j})}{|P| \cdot (|P|-1)},$$

where *Jaccard-Similarity-Index*$(A^\star_{q_i}, A_{q^\star_j}) = |A^\star_{q_i} \cap A^\star_{q_j}|/|A^\star_{q_i} \cup A^\star_{q_j}|$. The *average Jaccard similarity index* reflects diversity loss. High values of this measure, indicate low diversity.

3.2 Performance Analysis on the Training and Testing Sets

In order to evaluate the novel strategy based on *entropic recall*, we used the three proposed evaluation performance metrics: $\overline{\textit{Precision@10}}$, *Global-Recall*, and $\overline{\textit{Jaccard-Similarity-Index}}$. Figure 1 shows the evolution of these metrics during the training process for combinations **Co1** and **Co2**. It can be observed that although **Co1** achieves higher precision, the incorporation of the *entropic recall* metric in **Co2** clearly favors global recall and diversity. While $\overline{\textit{Precision@10}}$ and $\overline{\textit{Jaccard-Similarity-Index}}$ slightly increase over the generations, *Global-Recall* remains almost unchanged.

In order to determine if the evolved queries are effective when used on a new corpus we evaluated the $\overline{\textit{Precision@10}}$, *Global-Recall* and $\overline{\textit{Jaccard-Similarity-Index}}$ metrics on the test set. The questions addressed here are (1) whether the evolved queries are superior to the baseline queries (i.e., queries generated directly from the initial topic description), and (2) whether the queries evolved using the **Co2** strategy achieve a good performance on the test set.

Table 1 indicates that the queries from the last generation evaluated on the test set achieve a considerably higher precision than those queries generated directly from the topic description. A natural loss in diversity is observed for **Co1** and **Co2**. However, strategy **Co2** is better in attenuating this effect. Note that the evaluation metrics applied to the first generation of queries is independent of the evolutionary strategy applied.

4 Related Work

Information retrieval approaches based on EAs have manly focused on refining the initial set of results by means of improved queries [2, 1]. A comparative study of different MOEA strategies to learn boolean queries is presented in [16]. Other methods that apply EAs and MOEAs to evolve populations of topical queries are presented in [5, 6]. However, differently from the current study, these approaches do not address the problem of diversity preservation.

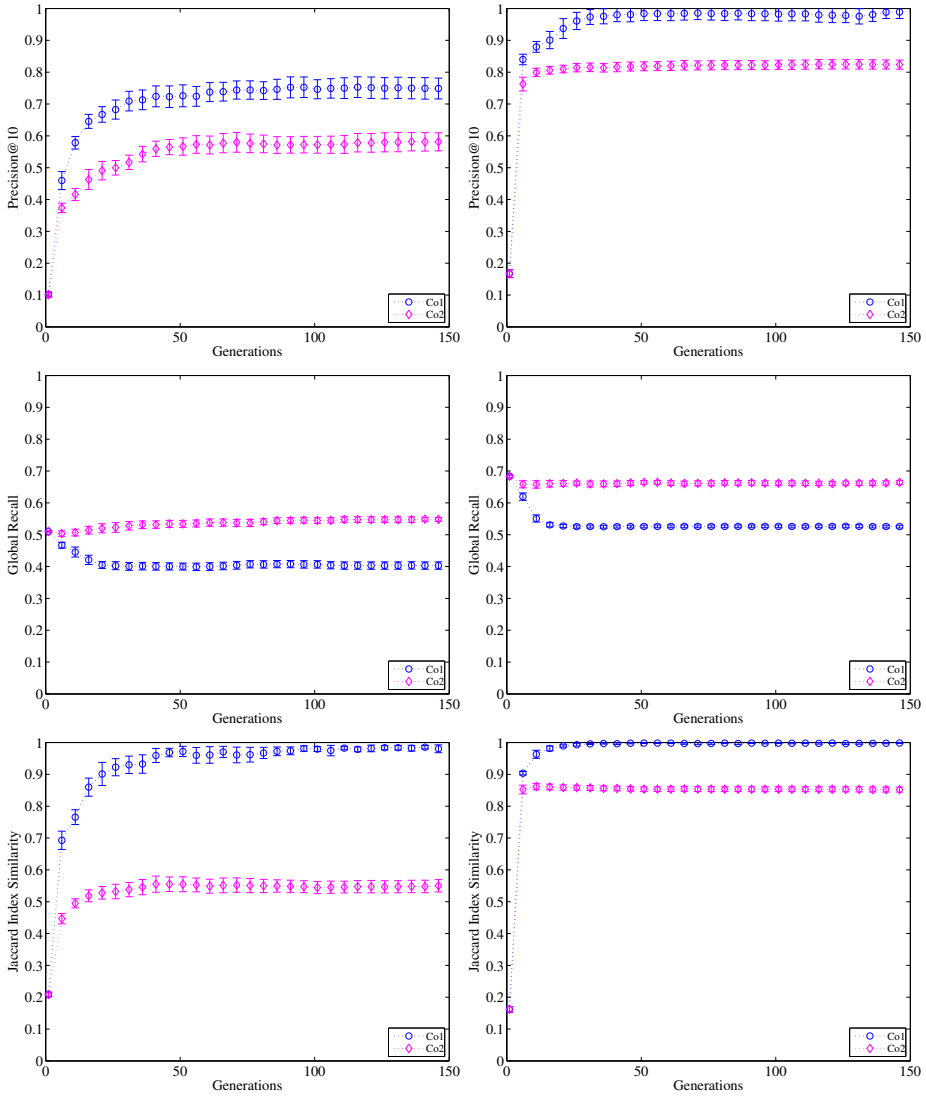

Fig. 1. Evolution of $\overline{Precision@10}$ (top), *Global-Recall* (middle) and $\overline{Jaccard\text{-}Similarity\text{-}Index}$ (bottom) for the topics BODY PAINTING (left) and AQUACULTURE (right).

Diversity in EAs has been addressed by niching methods such as crowding [9] and fitness sharing [12]. Another strategy usually exploited for diversity improvement in EAs is based on the concept of multiple populations [13], where the diversity of the entire population is maintained by independently evolving subpopulations.

The problem of diversification of search results has been mostly studied with the focus on optimizing the relevance and coverage of results at the document set level [4]. This is closely related to our goal of attaining coverage and diversity at the global level. However, instead of proposing strategies for attaining diversity, their goal is to compare

TESTING C.I.				
Topics	**Metrics**	**Co1 - Co2**	**Co1**	**Co2**
		First Generation	**Last Generation**	**Last Generation**
BODY PAINTING	$\overline{Precision@10}$	[0.058, 0.075]	[0.432, 0.509]	[0.342, 0.395]
	$GlobalRecall$	[0.482, 0.482]	[0.362, 0.388]	[0.480, 0.525]
	$\overline{Jaccard\text{-}Similarity\text{-}Index}$	[0.237, 0.272]	[0.975, 0.999]	[0.580, 0.662]
AQUACULTURE	$\overline{Precision@10}$	[0.158, 0.186]	[0.773, 0.845]	[0.735, 0.775]
	$GlobalRecall$	[0.696, 0.708]	[0.511, 0.529]	[0.672, 0.686]
	$\overline{Jaccard\text{-}Similarity\text{-}Index}$	[0.191, 0.210]	[0.985, 0.997]	[0.822, 0.835]

Table 1. Confidence intervals at the 95% level for the mean values of $\overline{Precision@10}$, *Global-Recall* and $\overline{Jaccard\text{-}Similarity\text{-}Index}$ evaluated on the test set using the first and the last generation of queries for the topics BODY PAINTING and AQUACULTURE.

clustering and diversification of search results using a unified evaluation framework, where the notion of *subtopic* plays a key role. The concept of "marginal relevance" introduced by [3] relates to our proposal in ranking results by considering their similarity to the previously retrieved results. However, these approaches focus on methods aimed at identifying diverse results associated with a single query rather than on generating diverse queries associated with a topic of interest, as is done by the methods proposed in this article. Query diversification mechanisms typically attempt to identify diverse queries that are semantically related to the original query [17, 15]. A comprehensive survey of search result diversification approaches is presented in [18]. It is worth mentioning that all of these approaches address the problem of diversification of search results by applying strategies considerably different from our proposal. In particular, none of these strategies rely on the application of EAs for topical retrieval.

5 Conclusions and Future Work

This paper proposes a novel strategy for evolving topical queries with special focus on diversity preservation. The main contribution is the formulation of an information theoretic fitness function that allows to enhance recall at the population level. This fitness function has been used in combination with precision at ten, resulting in a MOEA strategy that proved to be suitable in attaining good global recall, without a significant loss in accuracy.

As part of our future work we plan to carry out a more extensive analysis with a larger number of topics. In addition we plan to test new fitness functions aimed at preserving diversity at the query level. Another research direction includes the use of Genetic Programming to exploit the full potential of generalized boolean queries.

References

1. Pragati Bhatnagar and Narendra Pareek. Improving pseudo relevance feedback based query expansion using genetic fuzzy approach and semantic similarity notion. *Journal of Information Science*, 40(4):523–537, 2014.
2. Massimiliano Caramia, Giovanni Felici, and Alessia Pezzoli. Improving search results with data mining in a thematic search engine. *Computers & Operations Research*, 31(14):2387–2404, 2004.
3. Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336, New York, NY, USA, 1998. ACM.
4. Claudio Carpineto, Massimiliano D'Amico, and Giovanni Romano. Evaluating subtopic retrieval methods: Clustering versus diversification of search results. *Information Processing Management*, 48(2):358–373, March 2012.
5. Rocío L. Cecchini, Carlos M. Lorenzetti, Ana G. Maguitman, and Nélida B. Brignole. Using genetic algorithms to evolve a population of topical queries. *Information Processing and Management*, 44(6):1863–1878, 2008.
6. Rocío L. Cecchini, Carlos M. Lorenzetti, Ana G. Maguitman, and Nélida B. Brignole. Multiobjective Evolutionary Algorithms for Context-based Search. *Journal of the American Society for Information Science and Technology*, 61(6):1258–1274, 2010.
7. Raman Chandrasekar, Parikshit Sondhi, and Robert Rounthwaite. Topical search engines and query context models, November 3 2015. US Patent 9,177,045.
8. Carlos A. Coello Coello, Gary B. Lamont, and David A. Van Veldhuizen. *Evolutionary Algorithms for Solving Multi-Objective Problems*. Genetic and Evolutionary Computation. Springer-Verlag New York, Inc, New York, NY, 2nd edition, sep 2007.
9. Kenneth Alan De Jong. *An analysis of the behavior of a class of genetic adaptive systems*. PhD thesis, University of Michigan, Ann Arbor, MI, USA, 1975.
10. Kalyanmoy Deb. *Multi–Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Ltd., Chichester, W Sussex, UK, 2001.
11. Kalyanmoy Deb, Samir Agrawal, Amrit Pratap, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.
12. David E. Goldberg and Jon Richardson. Genetic algorithms with sharing for multimodal function optimization. In *Proc. of the 2nd Intl. Conf. on Genetic Algorithms on Genetic algorithms and their application*, pages 41–49, Hillsdale, NJ, USA, 1987.
13. Deepti Gupta and Shabina Ghafir. An overview of methods maintaining diversity in genetic algorithms. *IJETAE*, 2(5):56–60, 2012.
14. John H. Holland. *Adaptation in natural and artificial systems*. Bradford Series in Complex Adaptive Systems. The University of Michigan Press, Ann Arbor, MI, USA, 1975.
15. Youngho Kim and W. Bruce Croft. Diversifying query suggestions based on query documents. In *Proc. of the 37th Intl. ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 891–894, New York, NY, USA, 2014. ACM.
16. Antonio Gabriel López-Herrera, Enrique Herrera-Viedma, and Francisco Herrera. A study of the use of multi-objective evolutionary algorithms to learn boolean queries: A comparative study. *JASIST*, 60(6):1192–1207, 2009.
17. Hao Ma, Michael R. Lyu, and Irwin King. Diversifying query suggestion results. In Maria Fox and David Poole, editors, *Proc. of the 24th AAAI Conf. on Artif. Intell.*, pages 1399–1404. AAAI Press, July 2010.
18. Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. Search result diversification. *Foundations and Trends in Information Retrieval*, 9(1):1–90, March 2015.