Applied Soft Computing xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Topic relevance and diversity in information retrieval from large datasets: A multi-objective evolutionary algorithm approach

Rocío L. Cecchini, Carlos M. Lorenzetti*, Ana G. Maguitman, Ignacio Ponzoni

Instituto de Ciencias e Ingeniería de la Computación (ICIC), CONICET-UNS, San Andrés 800, Campus Palihue, Bahía Blanca, Argentina

ARTICLE INFO

Article history:
Received 30 April 2016
Received in revised form
24 September 2017
Accepted 10 November 2017
Available online xxx

Keywords:
Query reformulation
Information retrieval
Topic modeling
Diversity preservation

ABSTRACT

Enabling effective information search is an increasing problem, as technology enhances the ability to publish information rapidly, and large quantities of information are instantly available for retrieval. In this scenario, topical search is the process of searching for material that is relevant to a given topic. Multi-objective Evolutionary Algorithms have demonstrated great potential for addressing the topical search problem in very large datasets. In an evolutionary approach to topical search, a population of queries is automatically generated from a given topic, and the population of queries then evolves towards successively better candidate queries. Despite the promise of this approach, previous studies have revealed a common genotypic phenomenon: throughout evolution, the population tends to converge to almost identical sets of terms. This situation reduces the solution set to a few queries and leads to the exploration of a very limited region of the search space, which constitutes a limitation when users require different options from a topical search tool. This paper proposes and evaluates strategies to favor diversity in evolutionary topical search. These strategies rely on novel fitness functions, different parameterization for the crossover and mutation rates, and the use of multiple populations to favor diversity preservation. Experimental results conducted using these strategies in combination with the NSGA-II algorithm on a dataset consisting of more than 350,000 labeled web pages indicate that the proposed strategies show great promise for searching very large datasets, by helping to achieve query and search result diversity without giving up precision.

1. Introduction

Effective information retrieval from very large datasets has become increasingly important as technology enhances the ability to publish information rapidly and large quantities of information are instantly available for retrieval. In this scenario, topical search emerges as a useful procedure that allows seeking material related to a topic of interest [1,2].

There are many Big Data applications in which topical search has proven to be useful. For instance, topical search can be applied to support *task-based information search*, where an intelligent assistant monitors the user's activity and retrieves information by contextualizing user's information requirements [3,4]. Another application in which topical search can be usefully applied is *deep Web access*, where queries are formulated to harvest deep Web resources [5]. Topical search can also be applied to *satisfy persistent information interests*, where a software agent can play an intermediary role between a search interface and a user. In this kind of applications, the agent learns the user's interest profile and uses this profile to construct queries automatically and to filter results, offering information that is of interest to the user over an extended period of time [6–8]. Finally, topical search shows great promise in the area of *opinion mining*, where topical search is applied in social media to collect opinions about specific topics [9,10].

Several techniques exist to infer and represent a topic of interest to a user. This can be done by monitoring the user's activities, such as web pages or social web content being recently visited, documents being edited or emails being written [11–14]. Once the topic of interest has been represented, a new major challenge is to develop mechanisms to search for material related to the given topic.

By designing computational strategies for generating topical queries it becomes possible to automate the process of retrieving topic-relevant content. The development of such strategies will highly depend on assessing the effectiveness of the generated queries. This assessment is difficult because multiple objectives need to be considered to guide the query formulation process. In a broad sense, a query is effective if it retrieves the material the

* Corresponding author.
E-mail addresses: rlc@cs.uns.edu.ar (R.L. Cecchini), cml@cs.uns.edu.ar (C.M. Lorenzetti), agm@cs.uns.edu.ar (A.G. Maguitman), ip@cs.uns.edu.ar (I. Ponzoni).

https://doi.org/10.1016/j.asoc.2017.11.016
1568-4946/

user is looking for. Given a predefined set of relevant items for a topic of interest, classical performance metrics such as precision and recall can be computed for individual queries to evaluate topical relevance. The precision of a query is the fraction of documents retrieved (when posing the query) that are relevant, while the recall of a query is the fraction of relevant documents that are actually retrieved.

As discussed in recent studies, Evolutionary Algorithms (EAs) have proven to be successful in dealing with Big Data problems [15,16]. More specifically, the problem of topical search can be seen as a multi-objective optimization problem where the objective function to be maximized quantifies the effectiveness of a query [17,18].

In the topical-search optimization problem the initial set of candidate solutions is defined as the set of possible queries that can be presented to a search interface. Another aspect of this optimization problem is that the query space is a high-dimensional space, where each possible term in a query accounts for a new dimension. This kind of problems are computationally expensive and cannot be effectively solved using analytical methods. Besides, the problem of query optimization does not have optimal substructure, which means that an optimal solution cannot be constructed efficiently from optimal solutions to its subproblems [19]. Therefore, existing methods to solve complex problems by breaking them down into simpler steps are not effective for our purpose. On the other hand, a query can be considered effective even if it is not an optimal one, at the same time as multiple and diverse queries can provide satisfactory results. Therefore, we may be interested in finding many near-optimal queries rather than a single optimal one.

On the basis of the characteristics described above, EAs [20,21] are applicable to the problem of learning to automatically formulate optimal (or near-optimal) topical queries. Also, as mentioned above, more than one, possibly conflicting objectives can be defined to measure query effectiveness, resulting in a Multi-Objective Optimization Problem (MOOP) that may need to balance different criteria, such as precision, recall or other metrics.

In previous research work we have successfully applied Multi-Objective Evolutionary Algorithms (MOEAs) to evolve a population of topical queries [17,18], where the objectives to be maximized were precision and recall. However, evaluating individual queries in terms of precision and recall only provides a partial solution to the problem of measuring the performance of topical query generation. For the topical-search applications discussed earlier, additional desiderata for query generation strategies must be considered. In these applications, *the generated queries should be evaluated collectively rather than independently of each other*. Therefore, besides attempting to achieve high performance for the individual queries, *it is important to jointly attain high coverage and diversity while preserving local coherence*. In other words, we expect each topical query to be specific enough to return mostly relevant results, but the whole set of queries needs to be diverse enough not to miss many relevant results.

The problem of diversification of search results has been broadly explored by the information retrieval community (see [22] for a recent survey). In the meantime, population diversity has long been considered a critical issue in the performance of EAs [23]. However, to the best of the authors' knowledge, the solutions coming from the fields of information retrieval and evolutionary computation have been treated separately. Given the suitability of EAs to deal with the problem of topical search and the importance of diversification of search results, it is of great importance to investigate diversity preservation strategies in the context of evolutionary topical search. This article reviews and evaluates different aspects associated with the problem of diversity preservation in MOEAs for the specific scenario of topical search.

The main contributions of this work include:

- The definition of specialized fitness functions, where query performance not only depends on the results retrieved by the individual query but also depends on the results returned by other queries in the same population. The new metrics involve a reformulation of the classical precision and recall metrics, and we refer to them as *retrospective precision* and *retrospective recall*. To help achieve diversity, the new metrics penalize queries that retrieve results that have been previously obtained by other queries in the same population.
- The use of variable mutation and crossover rates by applying *hypo-crossover* (low crossover probability), *super-mutation* (high mutation probability), and *hyper-mutation* (very high mutation probability).
- The use of multiple populations of queries, to overcome the problem faced when a single population of queries converges to almost identical sets of terms.

The proposed strategies are evaluated using classical performance metrics as well as ad-hoc ones specifically aimed at assessing the diversity of the generated collection of queries as well as the coverage of relevant results retrieved by the entire population of queries. Population diversity is evaluated at the genotypic level, by assessing the diversity among the queries themselves, and at the phenotypic level using the Pareto front of the solutions. Finally, by analyzing the global recall reached by the population as a whole, it is possible to evaluate diversity at another phenotypic level.

2. Background and related work

2.1. Topical search

The process of searching for online data can be guided by diverse objectives. There are essential differences between searching for information to satisfy a particular consultation need and searching for resources to support the process of topical search. Usually, the purpose of a consultation is to find an answer to a particular question quickly and accurately. On the other hand, the purpose of topical search is to seek material based on a topic of interest. Different from the task of fulfilling a consultation need, topical search usually does not require a high-speed response. On the other hand, while accuracy is important for topical search, additional criteria should be considered such as high coverage and variety of results.

This article focuses on designing methods for topical search that can be used for applications where high speed is not crucial, and the main focus is to achieve high recall, novelty, and diversity without giving up precision. These applications can be built on top of existing search interfaces and can be used in different scenarios, such as searching for material to augment knowledge models [3,24], fulfilling long-term information needs [6–8,4], collecting resources for topical Web portals [25], accessing the deep Web [5], systematically reviewing medical digital repositories or records to identify research literature or previous experiences relevant to a given disease [26], among others.

The design of topic-based search methods usually requires the definition of methods for topic extraction and modeling, query generation, query refinement, and topic-based filtering. A great variety of methods have been proposed to search for material based on a topic or thematic context [11,27–29,18,30]. Typically, these methods represent the user thematic context as a set of terms and automatically generate queries based on this representation.

A seminal context-based search system is the *Remembrance Agent* [31], which operates inside the Emacs text editor and continuously monitors the user's work to find relevant text documents,

Please cite this article in press as: R.L. Cecchini, et al., Topic relevance and diversity in information retrieval from large datasets: A multi-objective evolutionary algorithm approach, Appl. Soft Comput. J. (2017), https://doi.org/10.1016/j.asoc.2017.11.016