Information Processing and Management 53 (2017) 454–472

Contents lists available at ScienceDirect

Information Processing and Management

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

Exploring the space of information retrieval term scoring functions

Parantapa Goswami*, Eric Gaussier, Massih-Reza Amini

Université Grenoble Alpes, CNRS, LIG/AMA, France

ARTICLE INFO

Article history:
Received 7 February 2016
Revised 7 November 2016
Accepted 13 November 2016

Keywords:
IR theory
Function generation
Automatic discovery
IR constraints

ABSTRACT

In this paper we are interested in finding good IR scoring functions by exploring the space of all possible IR functions. Earlier approaches to do so however only explore a small subpart of the space, with no control on which part is explored and which is not. We aim here at a more systematic exploration by first defining a grammar to generate possible IR functions up to a certain length (the length being related to the number of elements, variables and operations, involved in a function), and second by relying on IR heuristic constraints to prune the search space and filter out *bad* scoring functions. The obtained candidate scoring functions are tested on various standard IR collections and several simple but promising functions are identified. We perform extensive experiments to compare these functions with classical IR models. It is observed that these functions are yielding either better or comparable results. We also compare the performance of functions satisfying IR heuristic constraints and those which do not; the former set of functions clearly outperforms the latter, which shows the validity of IR heuristic constraints to design new IR models.

1. Introduction

The quest for new, high performing IR scoring functions has been one of the main goals of IR research, ever since the beginning of the field in the late forties. This quest has led to many IR models, from the boolean model and the vector space model (Salton & McGill, 1983) to more recent proposals as the language model (Ponte & Croft, 1998) and the relevance model (Lavrenko & Croft, 2003), BM25 (Robertson & Zaragoza, 2009) and more generally probabilistic models (Jones, Walker, & Robertson, 2000a, 2000b), the HMM model (Metzler & Croft, 2005), the divergence from randomness framework (Amati & Rijsbergen, 2002) with the information-based models (Clinchant & Gaussier, 2010), and learning to rank approaches (Liu, 2009).

These models originated from the fertile imagination and thinking of scientists, who either relied on first principles, within a given theoretical framework, to derive new scoring functions, or who devised learning procedures to identify the best function in a given family of functions, typically the family of linear functions. The space of possible IR scoring functions is however tremendously larger than the one explored through such a process. Quoting Fan, Gordon, and Pathak, 2004:

> *There is no guarantee that existing ranking functions are the best/optimal ones available. It seems likely that more powerful functions are yet to be discovered.*

* Corresponding author.
E-mail addresses: parantapa.goswami@imag.fr (P. Goswami), eric.gaussier@imag.fr (E. Gaussier), massih-reza.amini@imag.fr (M.-R. Amini).

http://dx.doi.org/10.1016/j.ipm.2016.11.003
0306-4573/

The motivation of finding the *best* or *optimal* IR scoring function has led researchers to automatically explore the space of IR functions, typically through genetic programming and genetic algorithms, which were seen as a way to automatically find IR functions by exploring parts of the solution space stochastically (Cummins & O'Riordan, 2006b; Gordon, 1988; Pathak, Gordon, & Fan, 2000). But such attempts have always been limited by the complexity of the search space and again only explored a portion of it, without a clear understanding on which parts were explored and which were not.

We follow here a different route that aims at a more systematic exploration of the space of IR functions. We do so by first defining a grammar to generate possible IR functions up to a certain length (the length being related to the number of elements, variables and operations, involved in a function), and second by relying on IR heuristic constraints (Fang, Tao, & Zhai, 2004) to prune the search space and filter out *bad* scoring functions. Such a possibility was mentioned in Cummins and O'Riordan (2006b) but had not been tried, to the best of our knowledge, before our first study presented in Goswami, Moura, Gaussier, Amini, and Maes (2014).

In addition, we perform extensive experiments on CLEF-3, TREC-3, TREC-5, TREC-6, TREC-7, TREC-8, WT10G and GOV2 to evaluate the performance of the scoring functions discovered by our search strategy. We show that these functions are simple yet effective on most of the collections and perform significantly better than other standard IR scoring functions as well as state-of-the-art genetic programming based approaches. While exploring the search space, we also compare the functions that do not satisfy the IR heuristic constraints with the ones that satisfy these constraints. As we will see, the latter set of functions significantly outperforms the former set, thus empirically validating the heuristic IR constraints used.

The current study builds upon the studies we presented in Goswami et al. (2014) and Goswami, Amini, and Gaussier (2015), and expands them in different ways: (a) we consider here functions of higher "length" (Section 4) so as to rely on a better and *deeper* exploration of the search space, (b) we provide here a detailed description of the framework used (Sections 4–6), (c) we show that the method is robust to the choice of the collection used to select the set of candidate scoring functions (Section 8.1), and (d) we show a complete comparison with genetic approaches (Section 8.6).

The remainder of the paper is organized as follows: Section 2 highlights the main contributions of this article; Section 3 discusses previous work and places our work with respect to the previous approaches; Section 4 introduces the function generation process and the grammar that underlies it; Section 5 describes how the function space can be formally pruned using IR heuristic constraints, while Section 6 describes the method followed to select the best performing functions from the pool of generated functions. Finally, Sections 7 and 8 describes the experiments conducted and the results obtained, while Section 9 concludes the paper.

2. Contributions

The key contributions of this article are summarized in the following points.

(a) *An automated discovery approach for systematic exploration of IR function space.* The primary contribution of this paper is the development of an automated discovery approach which can systematically explore the IR function space in order to find efficient IR scoring functions. For this, a context free grammar is defined that generates well-formed functions up to a certain length, and then intelligent strategies (such as heuristic IR constraints) are deployed to prune the set of candidate functions so as to focus on well-defined functions only. The paper contains a detailed explanation of the newly proposed framework supported by theoretical arguments (Sections 4 and 6). Extensive experiments are performed to analyze and validate different parts of the framework (Sections 7 and 8).

(b) *Empirical analysis of heuristic IR constraints.* Further, we provide an empirical validation of heuristic IR constraints in light of the proposed discovery approach (Section 8.2). Scoring functions satisfying IR constraints are compared against the functions which do not. Superior performance of the first set over the second justifies the usefulness of IR heuristic constraints. These results are also in line with other empirical studies that aimed to test the validity of these constraints. Also additional experiments are conducted to further analyze the impact of each constraint separately on the performance of the scoring functions.

(c) *Discovery of effective IR scoring functions.* This newly proposed approach enables the "discovery" of highly competitive scoring functions. Best among these discovered functions can almost always outperform standard IR models and the functions evolved using state-of-the-art genetic programming techniques. Based on our findings, we also propose, several scoring functions found by our approach that can be used either independently or along with standard IR models in four most common retrieval situations (Section 9).

3. Related work

The will to explore the space of scoring functions to discover interesting and new IR functions is not new. The first attempts were based on genetic algorithms (Goldberg, 1989) and genetic programming (Koza, 1992). Genetic algorithms are heuristic optimization strategies inspired by the principles of biological evolution. Starting with an initial population of solutions, referred to as individuals, genetic operations (as reproduction, mutation and crossover) are iteratively used to create new individuals that provide a better solution to the problem. Genetic programming is an extension of genetic algorithms where each individual is essentially a program (or an algorithm). We review these different approaches below.

The first applications of genetic algorithms in IR aimed at finding suitable document and query descriptions (Blair, 1990; Gordon, 1988). At each generation, new query descriptions are created from existing ones so as to improve the retrieval