Information Processing and Management 53 (2017) 1005–1025

Contents lists available at ScienceDirect

Information Processing and Management

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

Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems

David E. Losada[a,*], Javier Parapar[b], Alvaro Barreiro[b]

[a] *Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS), Universidade de Santiago de Compostela, Spain*
[b] *Information Retrieval Lab, Department of Computer Science, University of A Coruña, Spain*

ARTICLE INFO

Article history:
Received 31 October 2016
Revised 19 April 2017
Accepted 20 April 2017

Keywords:
Information retrieval
Evaluation
Pooling
Reinforcement learning
Multi-armed bandits

ABSTRACT

Evaluating Information Retrieval systems is crucial to making progress in search technologies. Evaluation is often based on assembling reference collections consisting of documents, queries and relevance judgments done by humans. In large-scale environments, exhaustively judging relevance becomes infeasible. Instead, only a pool of documents is judged for relevance. By selectively choosing documents from the pool we can optimize the number of judgments required to identify a given number of relevant documents. We argue that this iterative selection process can be naturally modeled as a reinforcement learning problem and propose innovative and formal adjudication methods based on multi-armed bandits. Casting document judging as a multi-armed bandit problem is not only theoretically appealing, but also leads to highly effective adjudication methods. Under this bandit allocation framework, we consider stationary and non-stationary models and propose seven new document adjudication methods (five stationary methods and two non-stationary variants). Our paper also reports a series of experiments performed to thoroughly compare our new methods against current adjudication methods. This comparative study includes existing methods designed for pooling-based evaluation and existing methods designed for metasearch. Our experiments show that our theoretically grounded adjudication methods can substantially minimize the assessment effort.

1. Introduction

To measure the effectiveness of Information Retrieval (IR) systems, it is customary to build benchmarks consisting of a document collection, a series of information needs, and a set of relevance judgments (Clough & Sanderson, 2013; Sanderson, 2010). Standard collections, like those developed under the Text Retrieval Conference (TREC) (Voorhees & Harman, 2005), are so large that exhaustively judging every query-document pair becomes infeasible. Furthermore, with complete relevance judgments, human assessors' time would be mostly dedicated to analyzing non-relevant documents. Such an exhaustive approach would be a waste of time and effort. The most productive use of human assessors occurs when they judge documents deemed to be relevant (Sanderson & Zobel, 2005). This is why most assessment processes in IR experimentation are supported by a sampling method called *pooling*.

Pooling is a traditional method (Sparck Jones & van Rijsbergen, 1975) that has been extensively used in campaigns like TREC, CLEF (Conference Labs of the Evaluation Forum), NTCIR (NII Testbeds and Community for Information Access Research)

[*] Corresponding author.
E-mail addresses: david.losada@usc.es (D.E. Losada), javierparapar@udc.es (J. Parapar), barreiro@udc.es (A. Barreiro).

http://dx.doi.org/10.1016/j.ipm.2017.04.005
0306-4573/© 2017 Elsevier Ltd. All rights reserved.

and INEX (Initiative for the Evaluation of XML Research). In pooled test collections, relevance judgments are only done for the documents that were among the top retrieved for some systems that participated in the evaluation campaign. The top retrieved documents have the greatest impact on effectiveness and, therefore, estimates made in this way are accurate.

The pooling methodology works as follows: i) given a document collection, the campaign's organizers define a search task that tests the ability of systems to retrieve relevant documents in response to a set of queries (often known as topics), ii) different groups participating in the task submit their system's results, iii) for each query, the top *k* documents from each participating system are pooled into a single set and presented to assessors for judging. Under this setting, the rankings of documents produced by the participating systems are commonly known as *runs; k* is the *pool depth*; and the resulting set of judgments are known as *qrels*. Having runs from a sufficient variety of systems and a reasonable setting for *k* (typically set to 100), the assessments can be done at an affordable cost and the resulting benchmark is solid and reusable (Voorhees & Harman, 2005).

If we can afford the cost of judging the whole pool then we can just pass the pooled documents to the assessors in random or arbitrary order. Instead, if our budget is limited, then we might want to judge a subset of the pool. This has motivated the emergence of a stream of proposals on how to adjudicate pooled documents for judgment (Cormack & Lynam, 2007; Cormack, Palmer, & Clarke, 1998; Moffat, Webber, & Zobel, 2007). An effective adjudication method selects documents from the pool following a given criterion. These adjudication methods, when compared with random or arbitrary alternatives, can substantially reduce the assessment effort required to produce a qrel file with a sufficient number of relevant documents (Moffat et al., 2007).

Selection strategies for labeling items from a pool of unlabeled items is of interest well beyond Information Retrieval. In many data mining applications, unlabeled data is abundant and manually labeling is expensive. Supervised learning –e.g. classification– requires a set of labeled examples and it is crucial to reduce the costs associated with creating the training data. This has motivated the emergence of a large number of studies on pool-based selection for learning (Reitmaier & Sick, 2013). Here, we are only concerned with the specifics of IR pooling but the lessons learned from our research are potentially applicable to other areas.

The assessment process needed to create an IR test collection can be seen as a "learning from interaction" process. The more assessed documents we have, the more we learn about the relative quality of the runs. Here we propose an innovative and formal adjudication approach based on multi-armed bandits. Research on document adjudication for IR evaluation has been mostly adhoc and has largely ignored the lessons learned in reinforcement learning.

The multi-armed bandit problem (Robbins, 1952), also known as K-armed bandit problem, is a long-established problem in reinforcement learning. Reinforcement learning is concerned with how an autonomous system interacts with an uncertain environment so as to maximize a numerical reward over some time period. The system is not explicitly told which actions to take but, instead, must learn which actions yield the most reward by testing them out. At any time point, each possible action has an estimated value and the system can opt to exploit its current knowledge (i.e. try the action whose estimated value is greatest). This exploitative choice is often called the greedy action. Alternatively, the system can choose to explore non-greedy actions. Exploring the environment in this way enables the system to improve its estimates for non-greedy actions. Exploitation is the right thing to do to maximize the short-term expected reward, but exploration may produce better results in the long run. Multi-armed bandits offer a theoretical framework for analyzing the trade-off between exploration and exploitation. Due to its generality, the exploration vs exploitation dilemma has been studied in many disciplines (Sutton & Barto, 1998), including medicine, economics, and ecology. For instance, balancing between exploration and exploitation has been employed to investigate the effects of different experimental treatments while minimizing patient losses (Press, 2009). We show here that balancing exploration and exploitation has also a practical application in IR evaluation and, in particular, in how to build qrels from a set of runs from different systems. Within this process, concentrating only on systems that currently look effective is risky. We can miss relevant documents that are only supplied by other apparently inferior runs. Multi-armed bandit algorithms are a natural solution to address this balance formally.

In summary, this paper contributes in the following interrelated aspects:

- We adapt multi-armed bandit models to address the problem of document adjudication in pooling-based evaluation of search algorithms. This innovative use of reinforcement learning leads to seven new effective adjudication methods that early identify relevant documents in the pools. Furthermore, this is a theoretically-grounded framework where we can analyze the exploration/exploitation dilemma. In doing so, we show how different document adjudication methods behave with respect to this dilemma.
- We conduct a thorough comparison of existing adjudication methods and confront their merits against effective models of metasearch. While a number of isolated studies have analyzed and proposed different adjudication methods, the literature is lacking a complete picture of their effectiveness. There is little experimental evidence on the relative merits of existing adjudication methods when compared with effective metasearch models. In this paper we try to fill this gap by performing a thorough evaluation of existing adjudication methods and a comparison of these methods against our seven bandit-based solutions. Our study comprises reference methods specifically designed for pooling-based evaluation and reference methods designed for metasearch. To the best of our knowledge, this is the first study that evaluates such a highly diversified portfolio of adjudication methods.
- We compare the most effective document adjudication methods with respect to their ability to identify relevant documents and in terms of the induced bias. By judging only a subset of the pooled documents we are inducing a bias with