Cost-effective construction of Information Retrieval test collections

David E. Losada
Centro Singular de Investigación
en Tecnoloxías da Información
(CiTIUS)
Universidade de Santiago de
Compostela
Santiago de Compostela, Spain
david.losada@usc.es

Javier Parapar
Information Retrieval Lab
Department of Computer Science
University of A Coruña
A Coruña, Spain
javierparapar@udc.es

Alvaro Barreiro
Information Retrieval Lab
Department of Computer Science
University of A Coruña
A Coruña, Spain
barreiro@udc.es

ABSTRACT

In this paper we describe our recent research on effective construction of Information Retrieval collections. Relevance assessments are a core component of test collections, but they are expensive to produce. For each test query, only a sample of documents in the corpus can be assessed for relevance. We discuss here a class of document adjudication methods that iteratively choose documents based on reinforcement learning. Given a pool of candidate documents supplied by multiple retrieval systems, the production of relevance assessments is modeled as a multi-armed bandit problem. These bandit-based algorithms identify relevant documents with minimal effort. One instance of these models has been adopted by NIST to build the test collection of the TREC 2017 common core track.

CCS CONCEPTS

• **Information systems** → **Test collections**; **Relevance assessment**; • **Computing methodologies** → *Reinforcement learning*;

KEYWORDS

Information Retrieval evaluation, relevance assessments, pooling, multi-armed bandits

ACM Reference Format:
David E. Losada, Javier Parapar, and Alvaro Barreiro. 2018. Cost-effective construction of Information Retrieval test collections. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Evaluation is essential for advancing in search technologies. With current document corpora, making deep assessments becomes infeasible. For each query, it is customary to extract a sample of documents that become candidates for assessment. This is the so-called *pooling* strategy [9]. Pooling, which has been adopted by most evaluation campaigns, consists of doing relevance judgments only for those documents that are retrieved at top positions by search systems participating in the creation of the collection. This pool of candidate documents is still large and it is often the case that we can only afford to judge a subset of the pool.

A number of strategies have been proposed to analyze the pool and adjudicate documents for relevance assessment. These adjudication strategies follow different intuitions and attempt to create robust *qrels* (query-relevance judgments) with minimal effort. A recent study [6] compared multiple pooling-based and meta-search strategies and found that some instances of multi-armed bandit models lead to formal and extremely effective adjudication strategies. These bandit-based models, which were originally proposed in [5], are briefly reviewed in the next section.

2 RELEVANCE ASSESSMENTS BASED ON MULTI-ARMED BANDITS

Under a pooling setting, we work with multiple retrieval systems whose role is to rank documents by estimated relevance. For each test query we start therefore with multiple rankings of documents (*runs*) and, with the assistance of human judges, we need to produce some relevance judgments. Initially, we have no knowledge on the quality of the runs but, as judgments come in, we gain knowledge about the relative effectiveness of the runs. This iterative assessment process can naturally be modeled with reinforcement learning.

K-armed bandits [8], or multi-armed bandits, are classic reinforcement learning methods that model the balance between exploitation (e.g., next judgment extracted from the best run so far) and exploration (e.g., next judgment from a *suboptimal* run). These models fit well with a pooling-based construction of qrels, where we need to trade between selecting the currently best system and selecting another inferior system that can eventually become a good supplier of relevant documents.

Robbins [8] defined the K-armed bandit problem as follows. A gambler is faced with K bandits (or slot machines). Each bandit has an unknown probability of giving a prize and, when played, supplies a (numerical) reward. The gambler chooses one bandit per round and attempts to maximize the overall reward. A key component of bandit-based algorithms is the *allocation* strategy, which is the element of these methods that decides which bandit to play next. Such bandit selection depends on past rewards and each allocation strategy follows a different approach to deal with the exploration versus exploitation dilemma.

Conference'17, July 2017, Washington, DC, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Pooling-based document adjudication can be addressed as a K-armed bandit problem where the runs are the bandits and playing a bandit consists of judging the top (unjudged) document from the run. In this way, if the top document supplied by the run is relevant then the run gets a positive reward. Otherwise, it gets a negative reward[1]. This outcome is used to update our counts about the relative quality of the runs and, next, we can move on to the next pick. In [5], we evaluated a number of well-known bandit allocation strategies, including random selection, ϵ_n-greedy, Upper Confidence Bound (UCB), and Bayesian Bandits (BB), and compared them with alternative pooling adjudication methods. This comparison not only gave new insights about the relative merits of existing methods, but also showed that some BB models are superior to state-of-the-art solutions. The study included stationary bandit models, which assume that the probabilities of the bandits do not change and treat all rewards equally, and non-stationary bandit models, where past and recent rewards can be assigned unequal weights.

In [6], we further compared the most effective bandit-based models with meta-search models, and expanded the study with a bias analysis. By judging a subset of the pool, document allocation methods induce a bias with respect to judging the entire pool. This bias is reduced as more documents are assessed. We thoroughly studied the bias induced by different pooling allocation methods. These experiments concluded that one simple instance of the BB models requires fewer relevance assessments than any other competing method and leads to a ranking of search systems that highly correlates with the official rankings (i.e. it has low bias).

When the systems participating in the evaluation campaign supply relevance scores (e.g. query-document similarity scores), we can try to inject this evidence into the process of selection of relevance judgments. This avenue of research was explored in [7]. We proposed effective rank fusion methods that model the distribution of retrieval scores supplied by the search systems. Our evaluation showed that this theoretically-grounded approach is competitive when compared to state of the art methods. Another contribution of this study was that we successfully included pseudo-relevance evidence into the estimation of the score distribution models.

3 TREC 2017 COMMON CORE TRACK

In 2017, NIST faced the challenge of building a new test collection for the traditional ad-hoc search task [1]. To meet this aim, the organizers of the common core track decided to test new collection construction methodologies that avoid the disadvantages of the classic depth-k pooling. Such classic strategy creates robust collections but does so at a high cost. The main desiderata was to construct a reusable test collection with minimal cost.

The organizers of this new TREC track considered the BB models proposed in [5, 6], evaluated them (in terms of the relevant documents found and in terms of bias) and found that these models are cost-effective and can rank runs much the same as the official (full pool) qrels. Following this evaluation, which was consistent with our previous studies, a specific instance of our BB models was implemented in the NIST servers and used to iteratively select documents for assessment (the server iteratively sent documents to the human assessors and received their relevance assessment). TREC is a world reference in search technology evaluation, and having one of our models as a working piece of the TREC pipeline is a really encouraging result of our research.

4 CONCLUSIONS

Relevance assessments are a bottleneck in the process of building an Information Retrieval collection. In this paper, we have reviewed recent studies that proposed cost-effective solutions for generating query-relevance judgments. Effectively building test collections is a need not only for well-known evaluation campaigns, such as TREC [10], or NTCIR [3], but also for research teams or companies that build their own testbeds, e.g. to evaluate vertical search or to support experimentation in specific domains [2, 4].

Furthermore, selection methods for labeling items from a pool of unlabeled items is of interest well beyond Information Retrieval. As a matter of fact, unlabeled data is pervasive in many data mining applications and it is crucial to have cost-effective ways to create training data.

ACKNOWLEDGMENTS

This work has received financial support from the i) "Ministerio de Economía y Competitividad" of the Government of Spain and FEDER Funds under the research project TIN2015-64282-R, ii) Xunta de Galicia (project GPC 2016/035), and iii) "Consellería de Cultura, Educación e Ordenación Universitaria" (accreditation 2016-2019, ED431G/08) and the European Regional Development Fund (ERDF).

REFERENCES

[1] J. Allan, D. Harman, E. Kanoulas, D. Li, C. Van Gysel, and E. Vorhees. 2017. TREC 2017 Common Core Track Overview. In *Proc. of the 26th Text Retrieval Conference.* 22387–22392.

[2] K. Balog and R. Neumayer. 2013. A Test Collection for Entity Search in DBpedia. In *Proc. ACM SIGIR (SIGIR '13).* 737–740.

[3] N. Kando, T. Sakai, and M. Sanderson (Eds.). 2016. *Proc. 12th NTCIR Conference on Evaluation of Information Access Technologies.*

[4] D. Losada and F. Crestani. 2016. A Test Collection for Research on Depression and Language Use. In *Proc. CLEF.* 28–39.

[5] D. Losada, J. Parapar, and A. Barreiro. 2016. Feeling Lucky? Multi-armed Bandits for Ordering Judgements in Pooling-based Evaluation. In *Proc. of the 31st ACM Symposium on Applied Computing (SAC '16).* ACM, 1027–1034.

[6] D. Losada, J. Parapar, and A. Barreiro. 2017. Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. *Information Processing & Management* 53, 5 (2017), 1005 – 1025. https://doi.org/10.1016/j.ipm.2017.04.005

[7] D. Losada, J. Parapar, and A. Barreiro. 2018. A rank fusion approach based on score distributions for prioritizing relevance assessments in information retrieval evaluation. *Information Fusion* 39 (2018), 56 – 71. https://doi.org/10.1016/j.inffus.2017.04.001

[8] H. Robbins. 1952. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* 58, 5 (1952), 527–535.

[9] K. Sparck Jones and C.J. van Rijsbergen. 1975. *Report on the need for and provision of an "ideal" information retrieval test collection.* Technical Report. University of Cambridge, Computer Laboratory.

[10] E. Voorhees and D. Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval.* The MIT Press.

[1] We worked with binary relevance and, thus, rewards were either 0 or 1.