Information Processing and Management 54 (2018) 37–59

Contents lists available at ScienceDirect

Information Processing and Management

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

Intelligent topic selection for low-cost information retrieval evaluation: A New perspective on deep vs. shallow judging

Mucahid Kutlu*,[a], Tamer Elsayed[a], Matthew Lease[b]

[a] Dept. of Computer Science and Engineering, Qatar University, Qatar
[b] School of Information, University of Texas at Austin, USA

A B S T R A C T

While *test collections* provide the cornerstone for Cranfield-based evaluation of information retrieval (IR) systems, it has become practically infeasible to rely on traditional *pooling* techniques to construct test collections at the scale of today's massive document collections (e.g., ClueWeb12's 700M+ Webpages). This has motivated a flurry of studies proposing more cost-effective yet reliable IR evaluation methods. In this paper, we propose a new *intelligent topic selection* method which reduces the number of search topics (and thereby costly human relevance judgments) needed for reliable IR evaluation. To rigorously assess our method, we integrate previously disparate lines of research on intelligent topic selection and *deep* vs. *shallow* judging (i.e., whether it is more cost-effective to collect many relevance judgments for a few topics or a few judgments for many topics). While prior work on intelligent topic selection has never been evaluated against shallow judging baselines, prior work on deep vs. shallow judging has largely argued for shallowed judging, but assuming random topic selection. We argue that for evaluating any topic selection method, ultimately one must ask whether it is actually useful to select topics, or should one simply perform shallow judging over many topics? In seeking a rigorous answer to this over-arching question, we conduct a comprehensive investigation over a set of relevant factors never previously studied together: 1) method of topic selection; 2) the effect of topic familiarity on human judging speed; and 3) how different topic generation processes (requiring varying human effort) impact (i) budget utilization and (ii) the resultant quality of judgments. Experiments on NIST TREC Robust 2003 and Robust 2004 test collections show that not only can we reliably evaluate IR systems with fewer topics, but also that: 1) when topics are intelligently selected, deep judging is often more cost-effective than shallow judging in evaluation reliability; and 2) topic familiarity and topic generation costs greatly impact the evaluation cost vs. reliability trade-off. Our findings challenge conventional wisdom in showing that deep judging is often preferable to shallow judging when topics are selected intelligently.

1. Introduction

Test collections provide the cornerstone for system-based evaluation of information retrieval (IR) algorithms in the Cranfield paradigm (Cleverdon, 1959). A test collection consists of: 1) a *collection* of documents to be searched; 2) a set of pre-defined user *search topics* (i.e., a set of topics for which some users would like to search for relevant information, along with a concise articulation of each topic as a *search query* suitable for input to an IR system); and 3) a set of human *relevance judgments* indicating the relevance of

* Corresponding author.
E-mail addresses: mucahidkutlu@qu.edu.qa (M. Kutlu), telsayed@qu.edu.qa (T. Elsayed), ml@utexas.edu (M. Lease).

http://dx.doi.org/10.1016/j.ipm.2017.09.002
Received 2 March 2017; Received in revised form 17 July 2017; Accepted 14 September 2017

collection documents to each search topic. Such a test collection allows empirical A/B testing of new search algorithms and community benchmarking, thus enabling continuing advancement in the development of more effective search algorithms. Because exhaustive judging of all documents in any realistic document collection is cost-prohibitive, traditionally the top-ranked documents from many systems are *pooled*, and only these top-ranked documents are judged. Assuming the *pool depth* is sufficiently large, the reliability of incomplete judging by pooling is well-established (Sanderson, 2010).

However, if insufficient documents are judged, evaluation findings could be compromised, e.g., by erroneously assuming unjudged documents are not relevant when many actually are relevant (Buckley, Dimmick, Soboroff, & Voorhees, 2006). The great problem today is that: 1) today's document collections are increasingly massive and ever-larger; and 2) realistic evaluation of search algorithms requires testing them at the scale of document collections to be searched in practice, so that evaluation findings in the lab carry-over to practical use. Unfortunately, larger collections naturally tend to contain many more relevant (and seemingly-relevant) documents, meaning human *relevance assessors* are needed to judge the relevance of ever-more documents for each search topic. As a result, evaluation costs have quickly become cost prohibitive with traditional pooling techniques (Sanderson, 2010). Consequently, *a key open challenge in IR is to devise new evaluation techniques to reduce evaluation cost while preserving evaluation reliability*. In other words, how can we best spend a limited IR evaluation budget?

A number of studies have investigated whether it is better to collect many relevance judgments for a few topics – i.e., *Narrow and Deep* (NaD) judging – or a few relevance judgments for many topics – i.e., *Wide and Shallow* (WaS) judging, for a given evaluation budget. For example, in the TREC Million Query Track (Carterette, Pavlu, Kanoulas, Aslam, & Allan, 2009), IR systems were run on ~ 10K queries sampled from two large query logs, and shallow judging was performed for a subset of topics for which a human assessor could ascribe some intent to the query such that a topic description could be back-fit and relevance determinations could be made. Intuitively, since people search for a wide variety of topics expressed using a wide variety of queries, it makes sense to evaluate systems across a similarly wide variety of search topics and queries. Empirically, large variance in search accuracy is often observed for the same system across different topics (Banks, Over, & Zhang, 1999), motivating use of many diverse topics for evaluation in order to achieve stable evaluation of systems. Prior studies have reported a fairly consistent finding that WaS judging tends to provide more stable evaluation for the same human effort vs. NaD judging (Bodoff & Li, 2007; Carterette & Smucker, 2007; Sanderson & Zobel, 2005). While this finding does not hold in all cases, exceptions have been fairly limited. For example, Carterette, Pavlu, Kanoulas, Aslam, and Allan 2008 achieve the same reliability using 250 topics with 20 judgments per topic (5000 judgments in total) as 600 topics with 10 judgments per topic (6000 judgments in total). A key observation we make in this work is noting that all prior studies comparing NaD vs. WaS judging assume that search topics are selected randomly.

Another direction of research has sought to carefully choose which search topics are included in a test collection (i.e., *intelligent topic selection*) so as to minimize the number of search topics needed for a stable evaluation. Since human relevance judgments must be collected for any topic included, using fewer topics directly reduces judging costs. NIST TREC test collections have traditionally used 50 search topics (manually selected from a larger initial set of candidates), following a simple, effective, but costly topic creation process which includes collecting initial judgments for each candidate topic and manual selection of final topics to keep (Voorhees, 2001). Buckley and Voorhees 2000 report that at least 25 topics are needed for stable evaluation, with 50 being better, while Zobel 1998 showed that one set of 25 topics predicted relative performance of systems fairly well on a different set of 25 topics. Guiver, Mizzaro, and Robertson 2009 conducted a systematic study showing that evaluating IR systems using the "right" subset of topics yields very similar results vs. evaluating systems over all topics. However, they did not propose a method to find such an effective topic subset in practice. Most recently, Hosseini, Cox, Milic-Frayling, Shokouhi, and Yilmaz 2012 proposed an iterative algorithm to find effective topic subsets, showing encouraging results. A key observation we make is that prior work on intelligent topic selection has not evaluated against shallow judging baselines, which tend to be the preferred strategy today for reducing IR evaluation cost. We argue that one must ask whether it is actually useful to select topics, or should one simply perform WaS judging over many topics?

Our Work. In this article, we propose a new *intelligent topic selection* method which reduces the number of search topics (and thereby costly human relevance judgments) needed for reliable IR evaluation. To rigorously assess our over-arching question of whether topic selection is actually useful in comparison to WaS judging approaches, we integrate previously disparate lines of research on intelligent topic selection and NaD vs. WaS judging. Specifically, we investigate a comprehensive set of relevant factors never previously considered together: 1) method of topic selection; 2) the effect of topic familiarity on human judging speed; and 3) how different topic generation processes (requiring varying human effort) impact (i) budget utilization and (ii) the resultant quality of judgments. We note that prior work on NaD vs. WaS judging has not considered cost ramifications of how judging depth impacts judging speed (i.e., assessors becoming faster at judging a particular topic as they become more familiar with it). Similarly, prior work on NaD vs. WaS judging has not considered topic construction time; WaS judging of many topics appears may be far less desirable if we account for traditional NIST TREC topic construction time (Voorhees, 2016). As such, our findings also further inform the broader debate on NaD vs. WaS judging assuming random topic selection.

We begin with our first research question **RQ-1**: *How can we select search topics that maximize evaluation validity given document rankings of multiple IR systems for each topic?* We propose a novel application of *learning-to-rank* (L2R) to topic selection. In particular, topics are selected iteratively via a greedy method which optimizes accurate ranking of systems (Section 4.3). We adopt MART (Friedman, 2001) as our L2R model, though our approach is largely agnostic and other L2R models might be used instead. We define and extract 63 features for this topic selection task which represent the interaction between topics and ranking of systems (Section 4.3.1). To train our model, we propose a method to automatically generate useful training data from existing test collections (Section 4.3.2). By relying only on pre-existing test collections for model training, we can construct a new test collection without any prior relevance judgments for it, rendering our approach more generalizable and useful. We evaluate our approach on NIST TREC