Summary of the Thesis: "Cheap IR Evaluation: Fewer Topics, No Relevance Judgements, and Crowdsourced Assessments"

Author: Kevin Roitero
Supervisor: Stefano Mizzaro
Reviewers: Ben Carterette and Guido Zuccon

Abstract of the Thesis (as written in the Thesis)

To evaluate Information Retrieval (IR) effectiveness, a possible approach is to use test collections, which are composed of a collection of documents, a set of description of information needs (called topics), and a set of relevant documents to each topic. Test collections are modelled in a competition scenario: for example, in the well known TREC initiative, participants run their own retrieval systems over a set of topics and they provide a ranked list of retrieved documents; some of the retrieved documents (usually the first ranked) constitute the so called pool, and their relevance is evaluated by human assessors; the document list is then used to compute effectiveness metrics and rank the participant systems. Private Web Search companies also run their in-house evaluation exercises; although the details are mostly unknown, and the aims are somehow different, the overall approach shares several issues with the test collection approach.

The aim of this work is to: (i) develop and improve some state-of-the-art work on the evaluation of IR effectiveness while saving resources, and (ii) propose a novel, more principled and engineered, overall approach to test collection based effectiveness evaluation.

In this thesis we focus on three main directions: the first part details the usage of few topics (i.e., information needs) in retrieval evaluation and shows an extensive study detailing the effect of using fewer topics for retrieval evaluation in terms of number of topics, topics subsets, and statistical power. The second part of this thesis discusses the evaluation without relevance judgements, reproducing, extending, and generalizing state-of-the-art methods and investigating their combinations by means of data fusion techniques and machine learning. Finally, the third part uses crowdsourcing to gather relevance labels, and in particular shows the effect of using fine grained judgement scales; furthermore, explores methods to transform judgements between different relevance scales.

Motivations and Aims of the Thesis

Effectiveness evaluation by means of test collections is not the only possible approach (user studies and log analysis, particularly in the case of companies, are also widely used), but its importance is indisputable and, perhaps, it is even what differentiates IR from related areas.

However, some limitations of such an approach can be identified. From a pragmatical viewpoint, it can be observed that the whole evaluation process is rather expensive, in terms of both human time and money. From a more general standpoint, one can notice that the literature is enormous; on the other hand, one can clearly feel that a lot of work seems more "artisanship" than engineering. We do not yet have an overall and complete understanding of what happens when the theoretical ideal evaluation setting is somehow degraded, as it is always the case in practice. This scenario makes one wonder if there is a more principled approach to address the evaluation problem. For example, it is particularly striking that in both test collection based initiatives and in-company private evaluation exercises, enormous amount of data are produced and call for a deeper relationship with the disciplines of data science, big data, and machine learning, that have much recently increased their importance — but such a relationship is nowhere in sight.

The thesis sets in the Information Retrieval field, precisely in the branch of research which investigates how to reduce the cost and the effort in the evaluation of Information Retrieval systems, in particular using test collections. Specifically, the thesis investigates about the reduction of the cost and the effort in the evaluation of Information Retrieval systems by means of three different approaches: the reduction of the topic set currently used (Part I), the evaluation performed with no human intervention (Part II), and the evaluation performed collecting crowdsourced relevance judgements (Part III). Summarizing, this thesis aims to reduce the effort of this whole process evaluation, preserving the benefits.

Summary of Contributions

Part I: On Using Fewer Topics in Information Retrieval Evaluation

Context

When evaluating the effectiveness of (IR) systems, the design of the measurement process has been examined by researchers from many 'angles': e.g. the consistency of relevance judgements, the means of minimizing judgements while maintaining measurement accuracy, and the best formula for measuring effectiveness. One aspect – the number and type of queries (*topics* in TREC terminology) needed in order to measure reliably – has been discussed less often. In general, there has been a trend in test collection construction of increasing the number of topics, but without much consideration of the benefits of such an approach.

Contributions

First, the thesis details a re-implementation of the software used to perform the state-of-the art experiments using a novel approach based on evolutionary algorithms. Then, it presents a successful attempt of the reproduction of the notable state-of-the-art results: [GMR09], [Rob11], and [BMR13]. Finally, it provides the generalization of such results to other effectiveness metrics and other TREC collections.

Then, the thesis presents a complete and exhaustive analysis on using fewer topics in the evaluation of retrieval systems. It shows that a larger ground truth topic set results in average and best subsets that are more highly correlated with the ground truth topic set than found in previous work. More in detail, as the cardinality of the ground truth increases, the size of the

subset (relative to ground truth) required to obtain a high correlation also decreases. Moreover, for large cardinalities, worst topic sets can be found that show very low correlations with ground truth. It provides a detailed analysis on the role of statistically significant differences among runs considering different topic sets. Finally, the analysis shows that an effective clustering techniques can be exploited to find more representative topics.

Part II: On Effectiveness Evaluation Without Relevance judgements

Context

Probably the most expensive part of building a test collection is to produce, for every topic, the relevance assessment for the documents retrieved by the retrieval systems participating in the competition. To reduce the effort of this process, it is common practice to pool a subset of the top 1000 documents retrieved by each system; the relevance assessment is then performed only for the pooled documents. Many researchers tried to reduce the effort of producing relevance assessment, in several different ways; a more extreme approach is to produce automatic relevance assessment, i.e., to evaluate the systems participating in a test collection initiative without any relevance judgements, in a completely automatic way [SNC01, WC03, Spo07]. In this part we focus on this approach.

Contributions

The thesis discusses discuss the reproduction the most important work on evaluation of retrieval systems performed in absence of relevance judgements. It presents many details useful for future reproducibility, presents the results in a uniform way, and generalizes such results to other test collections, evaluation metrics, and a shallow pool. Finally, it expands those results, obtaining two practical strategies that seem effective to, respectively, decrease the costs involved in test collection based evaluation.

Then, the thesis presents an extensive analysis over 17 prediction methods, 14 TREC collections, 15 accuracy measures, obtained by combining 3 effectiveness measures with 5 correlation measures, 4 data fusion approaches (plus variants), and 12 machine learning algorithms (plus variants) for the evaluation and combination of evaluation with out relevance judgement techniques. It provides a systematic and uniform analysis on individual method effectiveness across different collections, and previously unnoticed relationships between the individual methods. Furthermore, it shows that practical results on method combinations by means of machine learning algorithms can be exploited to provide a practical methodology for the researcher that wants to run an effectiveness evaluation without human relevance assessments.

Part III: On Crowdsourcing Relevance judgements and The Effect of The judgement Scale

Context

Over the last few years, the increasing size of document collections created the need to scale the gathering of relevance judgements. For this reason, crowdsourcing has become a consolidated methodology to create relevance labels for query-document pairs given a judgement pool. In order to produce crowdsourced relevance labels at a quality level comparable with that of expert

assessors a number of techniques have been proposed and evaluated in literature. A common approach is to collect relevance judgements for the same query-document pair from different crowd workers and to aggregate them together [AM12, HCM+12, VGK+14] thus allowing to remove noise in the labels. Past research also showed that asking for a justification for the judgements [MLKE16] can increase crowdsourced relevance judgement quality. In our work we leverage crowdsourcing to collect relevance judgements over different scales and build on top of existing crowdsourcing research in terms of quality checks and task design best practices.

Contributions

First, this thesis presents a systematic study comparing the effects of different relevance scales on IR evaluation. It shows many advantages of the fine grained (S100) scale as compared to coarse-grained scales like binary (S2) as well as a scale with four levels (S4), and to unbounded scales (ME). We show that S100 preserves many of the advantages of ME like, for example, allowing to gather relevance judgements that are much more fine-grained than the usual binary or 4-value scales. Assessors use the full spectrum, although sometimes with a preference for scores that are a multiple of ten. S100 has also demonstrated advantages over ME in terms of agreement with judgements collected on a binary and four level scales. Overall, our results show that S100 is an effective, robust, and usable scale to gather fine-grained relevance labels.

Finally, the thesis addresses the issue of transforming relevance scales. It shows that when reusing existing collections, it may be necessary to transform judgements that have been originally collected in a fine-grained scale into a different relevance scale.

References

[AM12] Omar Alonso and Stefano Mizzaro. Using crowdsourcing for TREC relevance assessment. *Inf. Process. Manage.*, 48(6):1053–1066, 2012.

[BMR13] Andrea Berto, Stefano Mizzaro, and Stephen Robertson. On using fewer topics in information retrieval evaluations. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, ICTIR '13, pages 9:30–9:37, New York, NY, USA, 2013. ACM.

[GMR09] John Guiver, Stefano Mizzaro, and Stephen Robertson. A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM Trans. Inf. Syst.*, 27(4):21:1–21:26, November 2009.

[HCM+12] Mehdi Hosseini, Ingemar J. Cox, Natasa Milic-Frayling, Gabriella Kazai, and Vishwa Vinay. On aggregating labels from multiple crowd workers to infer relevance of documents. In *Advances in Information Retrieval - 34th European Conference on IR Research, ECIR 2012, Barcelona, Spain, April 1-5, 2012. Proceedings*, pages 182–194, 2012.

[MLKE16] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. Why is that relevant? collecting annotator rationales for relevance judgments. In *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2016, 30 October - 3 November, 2016, Austin, Texas, USA.*, pages 139–148, 2016.

[Rob11] S. Robertson. On the Contributions of Topics to System Evaluation. In *Proc. ECIR*, ECIR 2011, pages 129–140, New York, NY, USA, 2011. Springer-Verlag New York, Inc. LNCS 6611.

[SNC01] Ian Soboroff, Charles Nicholas, and Patrick Cahan. Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 66–73, New York, NY, USA, 2001. ACM.

[Spo07] Anselm Spoerri. Using the structure of overlap between search results to rank retrieval systems without relevance judgments. *Information Provessing &Management.*, 43(4):1059–1070, 2007.

[VGK+14] Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 155–164, New York, NY, USA, 2014. ACM.

[WC03] Shengli Wu and Fabio Crestani. Methods for ranking information retrieval systems without relevance judgments. In *Proceedings of the 2003 ACM Symposium on Applied Computing*, SAC '03, pages 811–816, New York, NY, USA, 2003. ACM.