TREC: Topic engineeRing ExerCise

J Shane Culpepper Stefano Mizzaro* Mark Sanderson Falk Scholer
School of Computer Science and Information Technology, RMIT University
GPO Box 2476, Melbourne 3001, Victoria, Australia
{shane.culpepper,stefano.mizzaro,mark.sanderson,falk.scholer}@rmit.edu.au

ABSTRACT

In this work, we investigate approaches to engineer better topic sets in information retrieval test collections. By recasting the TREC evaluation exercise from one of building more effective systems to an exercise in building better topics, we present two possible approaches to quantify topic "goodness": *topic ease* and *topic set predictivity*. A novel interpretation of a well known result and a twofold analysis of data from several TREC editions lead to a result that has been neglected so far: both topic ease and topic set predictivity have changed significantly across the years, sometimes in a perhaps undesirable way.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation

Keywords

Evaluation, TREC, Topics

1. INTRODUCTION

Test collection-building conferences such as TREC, CLEF, or NTCIR aim at understanding and evaluating Information Retrieval (IR) system effectiveness, and are usually seen as exercises to develop better IR systems. We address a dual research question that to our knowledge has not been asked so far: Can these evaluation conferences be interpreted as a tool to develop better *topics* instead of better systems? Are the topics that have been produced over the years improving? We might even ask: When considering one of the best known of these conferences, could the acronym of TREC be recast from a system engineering exercise to a "Topic engineeRing ExerCise"?

*Permanent address: Dept. of Mathematics and Computer Science, University of Udine, Udine, Italy, mizzaro@uniud.it

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609531.

As part of an initial study, we present two possible approaches to quantify the "goodness" of a topic – *topic ease* and *topic predictivity*. Each of these approaches are explored in turn in Sects. 2 and 3. Sect. 4 summarizes our current progress, and outlines future work that our initial exploration of this area has generated.

2. TOPIC EASE

In the usual system-oriented approach, IR researchers try to understand if there is a trend in system effectiveness over multiple years using measures such as MAP.[1] Instead, in this section we examine whether topic ease has changed over time.

2.1 Background

Mizzaro and Robertson [4] define *easy* topics as "*topics to which runs tend to give high AP values*", and topic ease can be quantified using AAP.[2] However, this is not completely satisfying for our purposes: changes in the trend of AAP over the years of an evaluation exercise, such as TREC, could be caused by changes in the effectiveness of systems over those same years (a *system effect*). Fortunately there is a potential workaround, which we discuss next.

2.2 Experimental data

The workaround exploits a result that is now well-known in the TREC community: that ad hoc retrieval effectiveness on conventional TREC collections appears to have stabilized. Fig. 1 is adapted from Voorhees and Harman [7, Fig. 8] and derived from data published by Buckley and Walz [2, Table 3]: eight versions of the SMART IR system, developed over the first eight years of TREC, were run on the topics of eight editions of TREC, and their MAP was computed accordingly. This historical analysis of SMART was one of the primary reasons to discontinue the ad hoc track at TREC, the argument being that systems had reached a "plateau" in effectiveness gains, and that the effort of maintaining the track outweighed the knowledge being gained [6].

However, the data used to generate this graph can tell another story that has not been addressed as much. The SMART analysis takes topic variations into account implicitly (by running each system version on each topic set), but does not make them explicit. By focusing on topic variations, we can replot the same data in a dual way – for each version of the system, rather than for each set of topics – and produce Fig. 2. This shows that topic ease decreased over the first five years of TREC (i.e., in the years 1992–1996) and then increased in the last two. The trend over the eight SMART systems seems consistent.

[1]Mean Average Precision, the arithmetic mean of the average precision values for a system, or *run*, over all topics.
[2]Average Average Precision is the arithmetic mean of the Average Precision (AP) values for a topic, over all systems/runs.

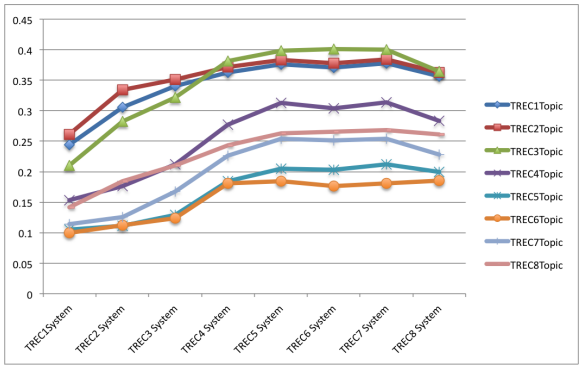

Figure 1: Effectiveness (MAP) of eight systems on the topics of eight TREC editions, adapted from Voorhees and Harman [7, Fig. 8].

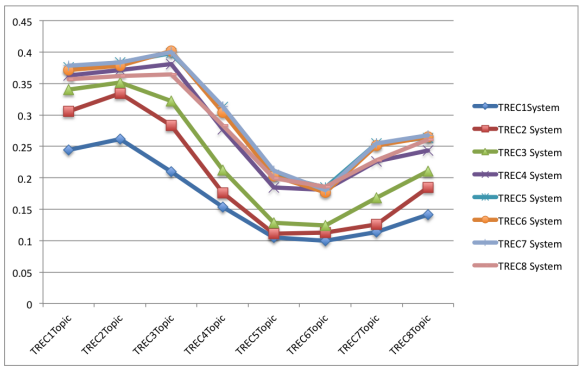

Figure 2: A topic-oriented representation of Fig. 1, showing how effectiveness (MAP) varies across the topic sets of eight years of TREC for eight fixed versions of the SMART system.

In Fig. 3 we overlay Fig. 2 with AAP values computed for all runs across TRECs 2–11.[3] Comparing the series in Fig. 3, one can see that despite quite different runs being measured, the AAP trend is consistent with the SMART trends: topics appear to get harder in the middle years of TREC ad hoc, easier in TRECs 7 and 8. Considering the later TRECs 9–11, the topics get harder again.

2.3 Discussion

While the data is old, we initially focus on the first eight years of TREC, since: (i) Figs. 1 and 2 can be drawn only for those years, and (ii) the goals of the tasks and the compositions of the document collections were largely the same.

The figures show that there is a substantial variation in topic ease across different years of TREC, with potentially undesirable effects. For example, a group participating in different TREC editions might be tempted to compare run effectiveness without taking topic ease changes into account. Also, the increased topic ease in TRECs 7 and 8 deserves particular attention since those are the TREC editions that have probably been used most frequently in data analysis and experiments (for example, it has been shown that to be effective in TREC 8, runs need to be effective on the easy topics [4]). As discussed at length by Mizzaro and Robertson [4], having harder topics is probably desirable as a track evolves: the

[3]We do not include TREC 1 data as it is not available on the TREC website, and we add TRECs 9–11, to be able to draw some general conclusions later.

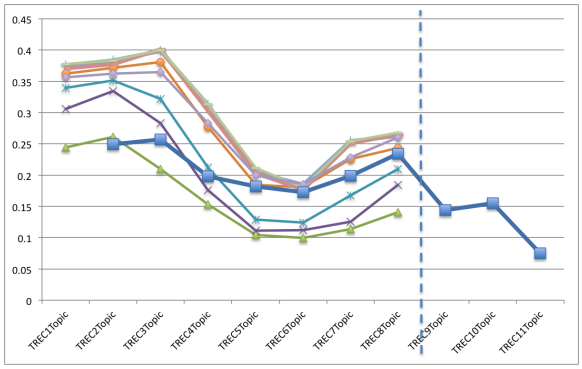

Figure 3: AAP values for TRECs 2–11

overall decreasing trends in Figs. 2 and 3 seems reassuring in this respect, although with important variations that are not completely understood, and with a notable drop in TRECs 9–11. More generally, being cognizant of changes in topic ease when analyzing trends in system effectiveness is important.

It is also possible that the trends we describe are attributable to other effects, which we consider here.

2.3.1 *System effects*

The AAP analysis might suffer from a *system effect*: perhaps topics are not becoming more difficult, but rather, the IR systems participating in TREC are on average becoming less effective. To consider this question, we plot the frequency of runs after they are assigned into one of five buckets of AAP values. The frequency values are computed over all runs submitted to the different years of TREC in Fig. 4. This graph shows that the number of low AAP values increased over time, and also around TRECs 5–6. However, this effect is likely due to topics becoming more difficult and not due to more poor systems being submitted to TREC. We draw this conclusion, because Fig. 2 shows that fixed versions of the SMART system become worse over different years of TREC.

2.3.2 *Topic variation*

The way in which topics were specified was itself an evolving process in the early years of TREC. For example, the TREC 1 and 2 topics contained a *concepts* field which was used effectively by several groups as a surrogate to the *summary* and *description* fields. From TREC 3 onward, the assessors who made relevance judgements were also the people who created the topics, and the concepts field was removed. As a result, groups began to depend on the *title* field for simple keyword queries. In TREC 4, only a shortened summary field was distributed to participants, greatly increasing the difficulty in identifying the key concepts automatically. Title and Narrative fields were reintroduced in TREC 5, and the basic topic format stabilized for the remaining ad hoc tracks.

However, these changes to topics do not appear to affect the trends seen in Fig. 2 and 3. The removal of concepts in TREC 3 did not uniformly impact SMART and AAP, and despite an improvement in topic detail from TREC 4 to TREC 5, both SMART and AAP were decreasing.

2.3.3 *Examining early web tracks*

In Fig. 3 and 4 we also show data from TRECs 9–11, which behaves differently from the first eight TRECs. We conjecture that this is largely due to a shift to web-based search tasks. In particular, TREC 11 has much lower AAP values; this is unlikely to be a factor of topics only, and indeed the TREC 11 task was different

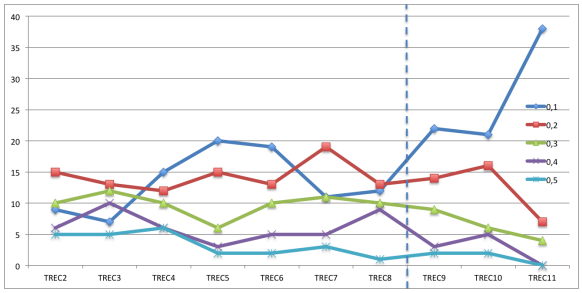

Figure 4: Trend of low and high AAP values

from the classic ad hoc tasks, attempting to introduce topic distillation. This turned out to be quite difficult initially, as there was confusion regarding exactly what the goals of the task were [6]. More generally, TRECs 9–11 were largely a transition period between the text-based ad hoc search tasks and HTML-based search tasks which could also leverage link information in the ranking algorithms. Historically, the web-based ad hoc tasks stabilized once GOV2 was introduced in 2004. We leave a thorough exploration of the trends of topic evolution in web documents to future work.

2.3.4 Conclusions on topic ease

From our preliminary analysis of past TREC runs, it would appear that examination of a system versus topic effect is worth considering. There is reasonable evidence that topics in TREC got harder and that this was not due to weaker systems being submitted or to changes in the way topics were defined. Analysis of such past data is difficult, but we consider there to be value in examination of these issues. The best way forward is for further work to disentangle the effect of topics and systems over the years.

3. TOPIC SET PREDICTIVITY

A second possible aspect of topic goodness is that a good topic should be a good predictor of overall system effectiveness.

3.1 Background

This interpretation is inspired by the work of Guiver et al. [3], Robertson [5] and Berto et al. [1]. In these papers, the authors attempted to understand whether the system evaluations carried out in TREC-like exercises could be accomplished using fewer than fifty topics. While no strong conclusions were drawn by the authors, byproducts of the work are useful here.

The first byproduct is the notion of *topic predictivity* (our term), i.e., the capability of a topic to predict overall system effectiveness. This can be rephrased as follows: if a system is evaluated using a subset of a test collection's topics, then in general the effectiveness scores using MAP (or any other metric) would be different. If that system was measured along with other IR systems using that topic subset, it would result in a different ranking of systems. The question then is how well the MAP (or system ranking) computed using the topic subset correlates with the MAP (or system ranking) computed using all the topics.

Previous work has shown that some topic subsets are more predictive than others, and the differences can be quite high [1, 3, 5]. Coming back to our research question, one might interpret "better topics" as "topics with higher predictivity" – we seek to understand if topic predictivity has changed over the years. The second byproduct that we use for our evaluations is the `BestSub` software [1, 3, 5], which can compute various correlations between effectiveness computed on a topic subset and on the full set of topics.

3.2 Experimental data

As with previous work [1, 3, 5], we use both linear correlation and Kendall's τ as measures of predictivity: the former measures how close MAP computed on a topic subset is to MAP computed on the full topic set. The latter measures how close a system rank measured on topic subsets is to a rank measured on full topic sets.

Fig. 5 (left) shows the best linear correlations (i.e., what we would find when selecting the possible best subset) that can be obtained when using topic subsets of cardinalities 3–12 over the TREC ad hoc test collections from TREC 2 to TREC 11.[4] Fig. 5 (right) shows similar data, but with Kendall's τ. Fig. 6 shows the same data for a random subset (i.e., what we would expect to find when sampling randomly from the population of topics; the average results over $10,000$ repeated samples are shown).

The trend in all figures is quite consistent across cardinalities and it is overall decreasing. Fitting linear models leads to regression line gradients that are negative in all cases. The slopes for both the linear and τ correlation "best" classes (Fig. 5) are statistically significant (t-test, $p < 0.05$) when at least four (eight) topics are used for the τ (linear) correlations.

3.3 Discussion

The overall decreasing trend means that the topic sets have become less predictive over the years, for the cardinalities evaluated here. The trend is particularly clear for Kendall's τ, indicating that the capability of topic subsets to predict the ranking of systems according to their effectiveness has fallen over time. At first sight, this is unlikely to be a desirable feature of topics. However, two corollaries should be considered.

First, the trend is more manifest for the best subsets than for the random subsets. This means that the theoretical potential predictivity (i.e., the predictivity that we would have if we were in some way able to select the best possible subset) is decreasing more than the practical effective predictivity (i.e., the predictivity that we might expect to find by selecting a random topic subset).

Second, the fact that subsets of topics appear to be becoming less able to predict system rankings of the full set of topics is not necessarily negative. If subsets of topics predict system rankings accurately, that tells us that there is some redundancy in the topics of old collections, and that is potentially a waste of effort. If newer collections do not have such redundancy, then this is positive: those collections include a more diverse set of topics, and are therefore presumably testing IR systems in a more complete way.

Coming back to the differences between TRECs 1–8 and TRECs 9–11, which are clear for topic ease (see Figs. 3 and 4), they become barely noticeable when considering topic predictivity (see Figs. 5 and 6). This means that whereas topics in TRECs 9–11 seem to be getting harder than in TRECs 1–8, that difference almost disappears when considering topic predictivity.

4. CONCLUSIONS AND FUTURE WORK

This paper proposes a dual approach to TREC data analysis: in place of trying to develop better IR systems, we try to understand how to develop better topics. While this work is a first step in this direction, there is much left for future work.

Although the trend in Fig. 1 is the same for each of the TREC topic sets (the eight lines), across the different versions of SMART

[4]The low cardinality topic sets (i.e., cardinalities 1–2) are noisy and prone to random effects; we do not take them into account.

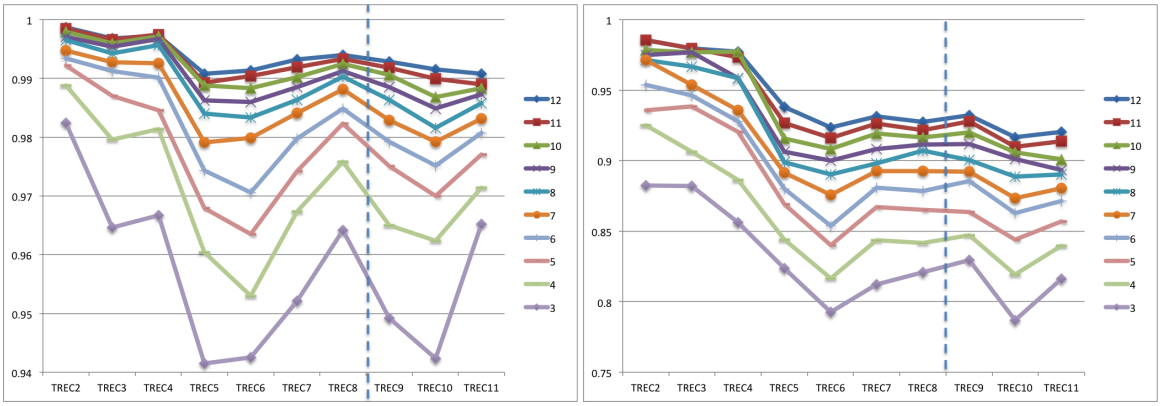

Figure 5: Topic set predictivity: best subsets, linear correlation (left) and Kendall's τ (right).

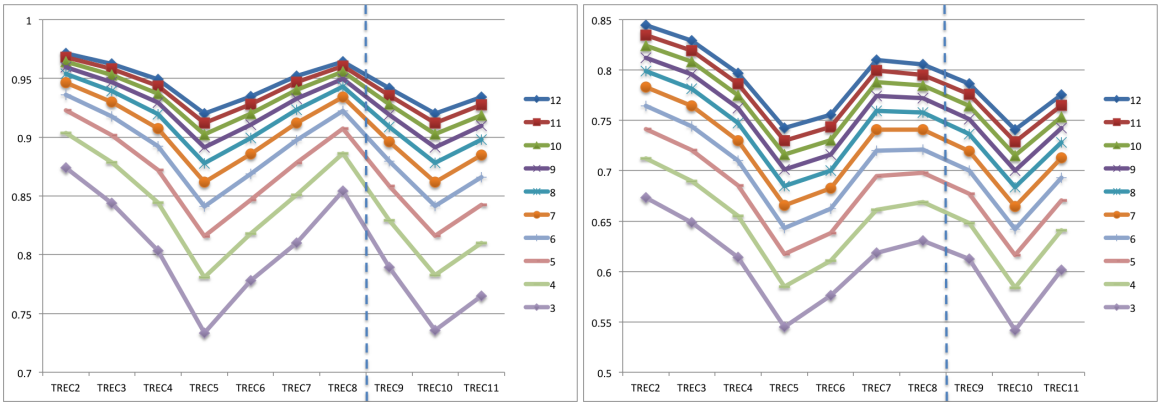

Figure 6: Topic set predictivity: random subsets, linear correlation (left) and Kendall's τ (right).

(the x-axis) the absolute magnitude of change (the y-axis) differs. Generally, the easier topics show larger variation, the harder topics show smaller. We will consider if statistical significance is more likely to be measured across topic sets with different topic ease.

In addition, there might be a "collection effect": over time the number of "possible" relevant results for each topic, on average, is increasing. Therefore, finding 1,000 documents in larger collections for each topic is easier now than in earlier collections. The role of the concept and title fields as a surrogate for a good keyword query is also not well understood. Note also that we ignore the shift from text to HTML in some of our analyses.

Other definitions of topic "goodness" can be proposed, and used to perform similar analyses. For example, a notion of "representativeness" of the topic space could be imagined for a topic set, although this would be difficult to quantify.

Our analysis is restricted to a limited number of TRECs, and should be extended to other collections, including other evaluation exercises such as NTCIR, INEX, FIRE, or CLEF. We have focused on MAP as the only effectiveness metric, but of course others can be used.

Finally, it would be interesting to repeat the work done by the SMART group in a broader fashion, i.e., with a set of systems working on the topics of all TREC tracks. Overall, we believe that there is a lot of work to be done, before the issue that we have started to address in this paper is fully understood.

Acknowledgments

This work was supported in part by the Australian Research Council (DP130104007) and also by a Google Faculty Research Award. Dr. Culpepper is the recipient of an ARC DECRA Research Fellowship (DE140100275).

References

[1] A. Berto, S. Mizzaro, and S. Robertson. On using fewer topics in information retrieval evaluations. In *ICTIR*, pages 30–37, 2013.

[2] C. Buckley and J. Walz. SMART in TREC8. In *TREC-8*, 2000.

[3] J. Guiver, S. Mizzaro, and S. Robertson. A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM TOIS*, 27(4):1–26, 2009.

[4] S. Mizzaro and S. Robertson. HITS hits TREC: exploring IR evaluation results with network analysis. In *SIGIR*, pages 479–486, 2007.

[5] S. Robertson. On the Contributions of Topics to System Evaluation. In *Advances in Information Retrieval*, volume 6611 of *LNCS*, pages 129–140, 2011.

[6] E. M. Voorhees and D. K. Harman, editors. *TREC: Experiment and evaluation in information retrieval*. MIT Press, London, 2005.

[7] E. M. Vorhees and D. K. Harman. Overview of the Eight Text REtrieval Conference (TREC-8). In *TREC-8*, 2000.