Easy Tasks Dominate Information Retrieval Evaluation Results

Thomas Mandl

Information Science
University of Hildesheim
Marienburger Platz 22, 31141 Hildesheim, Germany
mandl@uni-hildesheim.de

Abstract: The evaluation of information retrieval systems involves the creation of potential user needs for which systems try to find relevant documents. The difficulty of these topics differs greatly and final scores for systems are typically based on the mean average. However, the topics which are relatively easy to solve, have a much larger impact on the final system ranking than hard topics. This paper presents a novel methodology to measure that effect. The results of a large evaluation experiment with 100 topics from the Cross Language Evaluation Forum (CLEF) allow a split of the topics into four groups according to difficulty. The easy topics have a larger impact especially for multilingual retrieval. Nevertheless the internal test reliability as measured by Cronbach's Alpha is higher for more difficult topics. We can show how alternative, robust measures like the geometric average distribute the effect of the topics more evenly.

1 Introduction

Information Retrieval systems require the interaction between complex systems like weighting schemes and natural language processing. The resulting quality of a search system cannot be predicted because of this complexity and because components have a different effect for each language, document collection or usage context. Reliable evaluation is crucial on the other hand in order to optimize systems. The quality of information retrieval systems is most often evaluated based on a well accepted methodology [BV05]. Comparable conditions for systems are created by providing reference collections of text documents for which topics are constructed. These topics need to represent some potential user information need. The resulting documents of the competing systems are pooled and the documents in the pool are judged for relevance by human jurors. This method is called the Cranfield paradigm and forms the model for most of the scientific comparisons of search systems. In the last two decades, a large number of evaluation results have become available. They form the basis for research on the reliability of the Cranfield paradigm. The highest cost in the Cranfield evaluation paradigm is associated with the human effort to create relevance judgments. Most parameters within the model have been well researched over the last years.

Questions driving this research were: Are relevance judgments worth the resources dedicated to them? Do they lead to reliable system comparisons? Can fewer judgments also lead to the same results? The role of the human jurors has been explored by measuring their subjectivity. The interrater reliability can be shown to affect the absolute performance values but it only marginally modifies the ranking of the systems. Effort can be saved either by having fewer topics or by judging fewer documents per topic. Obviously, if more topics are developed, the reliability of the results is higher. Research needs to find an optimal balance between reliability of a test and the cost involved.

It has often been observed that the variance between topics is substantial. That means, both very easy and very hard topics typically occur in an evaluation set. The deviation between systems are mostly much smaller. Systems perform similarly whereas topics vary greatly. This observation has led to several approaches for improving systems. On one hand, systems should mainly optimize for the difficult topics which might have a larger impact on user satisfaction than easy topics. On the other hand, systems should try to guess which topics might be difficult and apply appropriate methods to them.

In this paper, we propose a new way of looking at evaluation results with regard to topic difficulty. This method is applied to the results of the Robust Task 2006 of the Cross Language Evaluation Forum (CLEF). We split the available topics into four groups according to their difficulty. The difficulty of a topic (search task) is understood as the mean or maximum average precision for this topic over several competing systems. This measure is not always intuitive and does typically not correlate with the number of relevant documents overall found for that topic.

The split proposed in this paper allows to measure the correlation of each subset with the results of the full set and to measure the internal test reliability of a test based solely on the subset. This analysis shows that the correlation between the full result and the ranking based on the easiest topics which are best solved by systems is usually higher than the correlation between the complete set and the harder topics. This is especially true for the multilingual sub-task.

2 Related Work: Evaluation Research

The evaluation of retrieval systems is directed towards a reliable ranking of systems in order to show which approaches and components work well for a particular retrieval task. The ranking of the systems is desired to be valid and reliable. The best system should also work well under real conditions.

The number of topics has often been an issue of discussion. Detailed analysis with small subsets of available retrieval results has led to the conclusion that 50 topics can produce a reliable result [BV02] and even 25 topics are sufficient [SZ05]. An approach to estimate how many topics are necessary is the leave-out technique. The topic set is subsequently reduced and the ranking of the systems is calculated based on the remaining subset. As long as the ranking remains stable and not too many position changes occur, one assumes that the reduced topic set would have also been sufficient.

Such an approach assumes that the ranking based on the complete set of topics represents the ground truth. We also take this approach in the analysis presented in the following section. However, it needs to be noted that this goodness measure for topics is a problematic assumption. Presumably, each topic tests some features of the system. These features may be how well a systems handles frequent or infrequent words, how it deals with topics with few or many hits, how it handles named entities or how well it deals with specific issues of stemming. Each of these capabilities may be relevant during the real use of the system and therefore each topic could be considered as a valid test for a system. In practice, it is of course unknown which features a topic would test. It typically is the case that some topics agree more on the ranking of the systems than others. Some correlate more than others and might be redundant. Network analysis techniques have been applied to identify the authoritativeness of a topic [MR07]. Identifying an optimal topic set would require more knowledge on the usage scenario as well as the type and frequency of topics during real use.

There is also research which calls for more topics but a smaller pool which contains much fewer documents [CAS07]. Such a shallow pool requires less human effort for relevance assessment, nevertheless, increasing the number of topics is supposed to boost reliability. The human effort could even be further decreased if it is directed toward topics and documents which allow a better ranking of systems during the assessment [MWZ07]. This idea is based on the observation that it would not make sense to judge a document which all systems have retrieved because this document does not discriminate between systems. Further research is necessary to investigate if that is the case for many evaluation settings. Many different performance measures have been suggested for information retrieval. In recent years, binary preference (BPref) has been widely adopted as a metric for test designs where only a small portion of the documents can be assessed [BV04]. Often, alternative measures correlate highly. Then one might argue that new measures are not necessary. However, when the results in the system rankings differ strongly, then these metrics measure different aspects of retrieval systems [Ro06]. It is not yet well understood what these aspects are and in which cases the use of a variety of measures makes sense.

A measure for the reliability of information retrieval tests has been suggested from the perspective of social science test theory [BL07]. The measure Cronbach's Alpha is a variance based value which uses the results of all base experiments for calculating the overall reliability of a test. Values above 0.8 are typically considered as indicators for reliable tests and indicate that the test is internally consistent and does not measure several factors. The values for various experiments within the Text Retrieval Conference (TREC) lie between 0.85 and 0.93 [BL07]. For a retrieval test, Cronbach's Alpha relates the sum of the variance of the systems with the variance of the AP sum for each topic. It will also be used in the following sections.

Retrieval systems as well as evaluation measures are desired to be robust. Robustness can have many meanings in everyday life. A general definition is the following one: "Robust … means … capable of functioning correctly, (or at the very minimum, not failing catastrophically) under a great many conditions" (reference.com).

For an information retrieval system there are many conditions or contexts of use. Consequently, robust IR means the capability of an IR system to work well (and reach at least a minimal performance) under a variety of conditions (topics, difficulty, collections, users, languages …). The evaluation of robust retrieval has been motivated by the fact that the variance for topics has been very large even for top performing systems. Even these good systems achieve only poor results for some topics. Improving on these topics would greatly enhance their overall quality as perceived by the user. Users remember poor performance often better than excellent performance. It is important to "ensure that all topics obtain minimum effectiveness levels" [Vo05]. In the special case that the average precision for one topic is 0, a value of 0.0001 is assumed.

As pointed out by Robertson, the geometric mean can also be considered as the mean of the logarithms of the basic results [Ro06]. Instead of the logarithm, another non-linear transformation could be used to stress the importance of the difficult topics. At the moment, the decision is arbitrary. In the long term, the transformation should be based on the observation of user behavior.

In an analysis with the smaller subsets of topics and a correlation analysis with the overall result of multilingual retrieval runs, it was shown that robust evaluation measures lead to different results than traditional measures [MW08]. Further analysis of the topic difficulty and its influence on the overall retrieval results are necessary.

In this paper, we propose and use a method for analyzing retrieval results in order to detect the relation between the difficulty of a topic and its influence on evaluation results. The same method is applied for a scenario with a robust evaluation measure. That way we can show that the geometric mean is a better aggregating measure especially for multilingual retrieval.

3 Context: Robust Task at the Cross Language Evaluation Forum

An evaluation track for robust retrieval has been established at the Text Retrieval Conference (TREC). This track does not only measure the average precision over all queries but also emphasizes the performance of the systems for difficult queries. In order to perform well in this track is more important for the systems to retrieve at least a few documents for difficult queries than to improve the performance in average [Vo05].

The robust task is user oriented because users often remember bad topics better than positive experiences. In order to allow a system evaluation based on robustness, more queries than for a normal ad-hoc track are necessary. The concept of robustness was extended in TREC 2005. Systems need to perform well over different tasks [Vo05].

The Cross Language Evaluation Forum (CLEF, www.clef-campaign.org) is a large European evaluation initiative which is dedicated to cross-language retrieval for European languages [Ma08, AN09]. A robust task has been organized for the first time at CLEF 2006 [Ma06].

The robust task uses test collections previously developed at CLEF. These collections contain data in six languages (Dutch, English, German, French, Italian and Spanish) and were almost constantly used at CLEF 2001, CLEF 2002 and CLEF 2003. There are approximately 1.35 million documents and 3.6 gigabytes of text in the collection. The robust task at CLEF 2006 received 133 runs by eight groups for ten sub-tasks. These tasks included monolingual retrieval as well as cross-lingual retrieval where query and document language are different from each other [Nu07, MW08]. For the multilingual task, the participants chose one query language and need to search all document collections. The result set needs to rank documents in all languages in one list. The usage scenario is a searcher who can read in a foreign language but who can form a more efficient query in his mother tongue.

4 Method

The following methodology was applied. From the robust task, all sub-tasks with at least ten submitted experiments (runs) were selected for the analysis. For the bilingual tasks, not enough experiments were submitted. If more than ten runs were available for a task, only the best ten were included in the analysis. Most sub-tasks contain some runs which need to be considered as negative outliers. This happens when research groups accidentally do not index the full collection or try some very different approaches for their experiments. These outliers are eliminated. The goal of retrieval experiments is to find the best system and not the perfect ordering for weak experiments. For all these runs, the performance of each system for each topic was known. Each sub-task is based on a different document collection and a translated version of the same topics.

In the next step, the topics were ordered according to their difficulty. For that, we adopted the most often used definition of topic difficulty: the mean average performance of all systems for that one topic [EGK02, Kw05]. Other measures like maximum and minimum performance for that topic and variance over systems for the topic as a potential measure for room for improvement were tried. The topic orderings based on these measures did not lead to consistent results. In future studies, other orderings based on non-linear transformations favoring the systems performing best for that topic could be tested.

The correlations between these topic orderings for individual sub-tasks are very low and not even weak. The largest absolute value is a correlation of 0.23 between multilingual and monolingual English. Most other values lie below 0.1. This confirms previous findings that topics are not difficult inherently but only in connection with a particular collection. After the ordering of the topics, four groups of topics based on difficulty were formed. For each of these quartiles of topics, a ranking of the systems was established. Because the robust task comprises 100 topics, an analysis with four sub-groups of topics is possible. The split leads to a size of subsets of topics which still has a considerably high reliability. This will be shown by the calculation of the internal test reliability.

The rankings of the systems were subsequently compared. They were determined for each of the four groups once by the mean average of the results for the 25 topics and once by the geometric mean. The rankings within the group were then compared to the ranking based on the full set by the mean average and the geometric average respectively. We also were interested in the correlation between the mean average (MAP) and the geometric average (GMAP) for the full topic sets.

Finally, the internal test reliability was calculated by Cronbach's Alpha measure. This was done for the full and the partial topic sets. It should be noted that there is only one Cronbach's Alpha and not one for MAP and one for GMAP. Cronbach's Alpha is based on the individual observations and MAP and GMAP can be seen as different ways to reach one quality measure through different aggregation of the individual observations.

5 Results and Discussion

The main observation from the resulting ranking comparison is that the easiest topics correlate better with the full ranking than harder topics. This is especially obvious for the multilingual sub-task. As figure 1 shows, the correlation is lowest for the hardest topics and higher for each following quartile of topics. It is obvious, that the easy topics have a higher influence on the final result or they have a higher goodness in correlating well with the full result. In a situation where the topic difficulty would be known before the evaluation and a reduction of the set would be necessary due to resource limitations, only easy topics would be selected. The easy tasks would determine the evaluation results. This seems to be negative for the evaluation as a whole. Hard topics are more likely to have a high impact on user satisfaction. Bad results lead to unsatisfied users. From the system developer's point of view, system failures for hard topics can give more insight into weaknesses of systems [HB04].

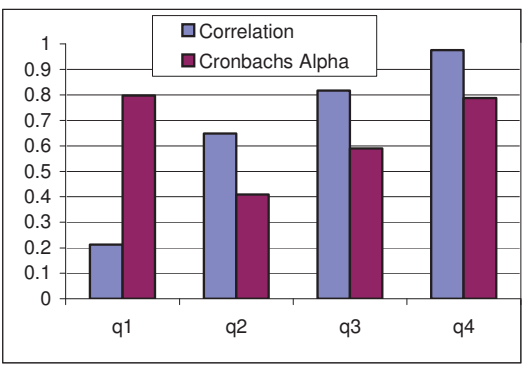

Figure 1: Correlation and Cronbach's Alpha for the Multilingual sub-task

As mentioned in the state-of-the-art overview above, robust measures can capture this better than the mean average of all topics. Figure 5 illustrates this effect. When the geometric mean (GMAP) is used to aggregate the results, the correlation between the subsets and the full set of topics is much more evenly distributed over the four quartiles. In addition, the correlation strength is lower for each set. The proposed analysis method illustrates the strength of the effect and the advantage of using the geometric mean of the individual observations.

The phenomenon observed for the multilingual sub-task cannot be seen as clear for the monolingual tasks which were included in the analysis. Figures 2, 3 and 4 show that the easiest quartile exhibits at least the second strongest correlation over all sub-tasks. The highest correlation always occurs in one quartile in the easier half. This shows that easier topics also earn a higher goodness for these tasks.

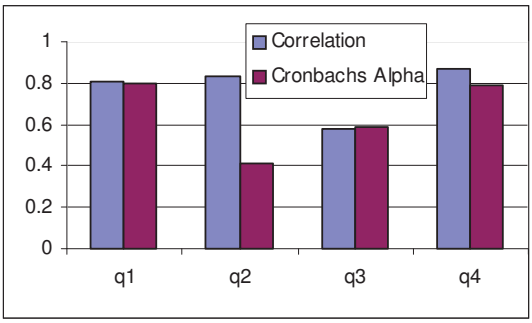

Figure 2: Correlation and Cronbach's Alpha for the Monolingual English sub-task

At the same time we can observe another interesting trend. The subset with the hardest and the easiest topics always have the best internal test reliability. The topics in the middle range are not so well suited to discriminate between systems. This finding needs to be considered in the context of the previous observation that hard topics do not deserve enough attention. At the same time they are good in discriminating between systems. Obviously there is a difference between the sub-tasks concerning the variances between the individual observations.

Why do the subsets for the mono- and multilingual tasks exhibit different phenomena? Cross-lingual retrieval requires an additional translation step. Typically, the query is automatically translated into the document languages before retrieval is initiated. This leads to lower performance. Due to the errors associated with machine translation, a decrease of about 10% to 20% in MAP is typically observed. That means, for the most often used definition of topic difficulty there are more hard topics in the multilingual retrieval set. This difference in the distribution of the average precision values for the topics is shown in figure 6. These topics might not be hard inherently but due to the necessary translation. When there are more hard topics in the set, the aggregation method is of more importance. Mean and geometric average for the multilingual task lead to more diverging results than for the monolingual tasks.

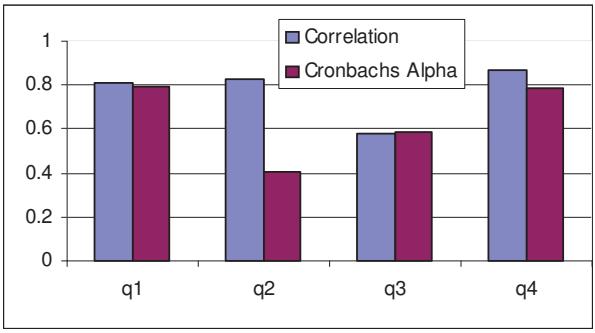

Figure 3: Correlation and Cronbach's Alpha for the Monolingual Spanish sub-task

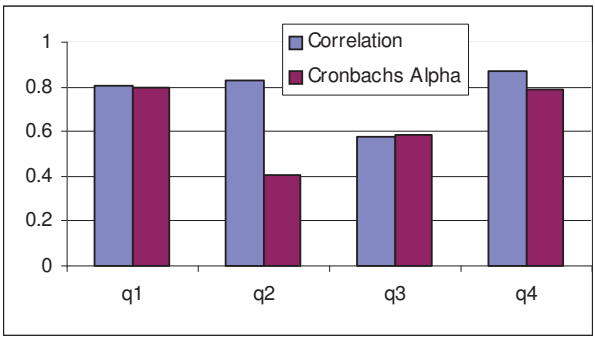

Figure 4: Correlation and Cronbach's Alpha for the Monolingual French sub-task

6 Conclusion and Future Work

The analysis presented here hints that an evaluation of multilingual retrieval focusing on robustness leads to substantially different results than standard evaluation measures. The analysis showed that the topic difficulty in cross-lingual retrieval is distributed in a manner that robust measures like the GMAP need to be considered. The influence of a topic set is measured as the correlation with the full set in the proposed method. In future studies, alternative measures should also be considered.

The current robust task (at CLEF 2008 and 2009) analyzes the effect of word sense disambiguation data on the robustness. For that end, annotated collections and topics are provided. Systems can extract the knowledge that a word is ambiguous and what meaning it has in each sentence. The effect of this information on the retrieval quality is currently being analyzed [AN09].

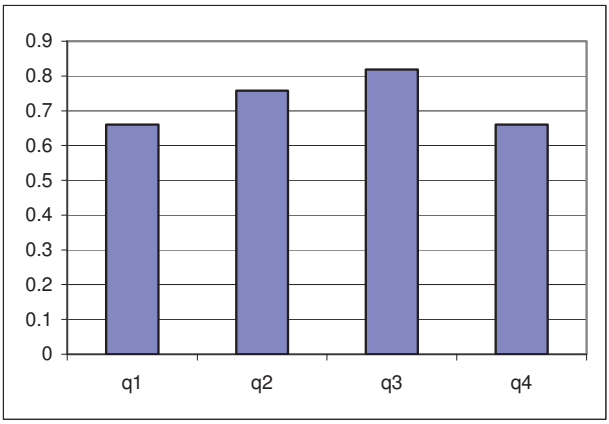

Figure 5: Multilingual Task: Correlation of the GMAP measure between small sets and full set

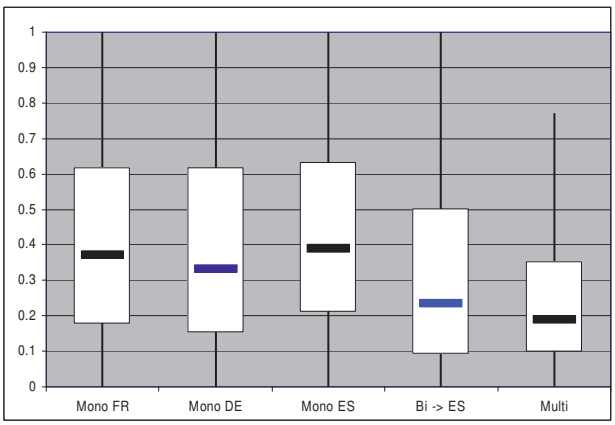

Figure 6: Distribution of Topic Difficulty for Several Tasks [MW08]

References

[AN09] Agirre, E.; Di Nunzio, G.; Ferro, N.; Mandl, T.; Peters, C.: CLEF 2008: Ad Hoc Track Overview. In: Evaluating Systems for Multilingual and Multimodal Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, Revised Selected Papers. Berlin et al.: Springer [LNCS]. *To appear.*

[BL07] Bodoff, D.; Li, P.: Test Theory for Assessing IR Test Collections. In: Proc Annual Intl. ACM Conf. on Research and Development in Information Retrieval (SIGIR) Amsterdam. 2007. 367-374.

[BV02] Buckley, C.; Voorhees, E.: The Effect of Topic Set Size on Retrieval Experiment Error. In: Proc Annual Intl. ACM Conf. on Research and Development in Information Retrieval (SIGIR) Tampere, 2002. 316-323.

[BV04] Buckley, C.; Voorhees, E. M.: Retrieval Evaluation with Incomplete Information. In: Proc Annual Intl ACM Conf. on Research and Development in Information Retrieval (SIGIR) 2004. 25-32

[BV05] Buckley, C.; Voorhees, E.: Retrieval System Evaluation. In: TREC: Experiment and Evaluation in Information Retrieval. Cambridge & London: MIT Press. 2005. 53-75.

[CAS07] Carterette, B.; Allan, J.; Sitaraman, R.: Minimal test collections for retrieval evaluation. In: Proc 29th Annual Intl. ACM Conf. on Research and Development in Information Retrieval (SIGIR) Seattle. 2006. 33-40.

[EGK02] Eguchi, K., Kando, N. and Kuriyama, K.: Sensitivity of IR Systems Evaluation to Topic Difficulty. In: Proc. of the Third International Conference on Language Resources and Evaluation (LREC 2002) 585-589.

[HB04] Harman, D.; Buckley, C.: The NRRC reliable information access (RIA) workshop. Proc 27th Annual Intl. Conf. on Research and development in information retrieval (SIGIR) 2004. 528-529.

[Kw05] Kwok, K.: An Attempt to Identify Weakest and Strongest Queries. In: SIGIR Workshop Predicting Query Difficulty. 2005. http://www.haifa.il.ibm.com/sigir05-qp

[Ma06] Mandl, T.: Benutzerorientierte Bewertungsmaßstäbe für Information Retrieval Systeme: Der Robust Task bei CLEF 2006. In: Effektive Information Retrieval Verfahren in Theorie und Praxis: Proc d. Fünften Hildesheimer Evaluierungs- und Retrievalworkshop (HIER 2006) http://web1.bib.uni-hildesheim.de/edocs/2006/519937899/meta/

[Ma08] Mandl, T: Recent Developments in the Evaluation of Information Retrieval Systems: Moving Toward Diversity and Practical Applications. In: Informatica – An International Journal of Computing and Informatics vol. 32. 2008. pp. 27-38.

[MW08] Mandl, T.; Womser-Hacker, C.; Ferro, N.; Di Nunzio, G.: How Robust are Multilingual Information Retrieval Systems? In: Proc. ACM SAC Symposium on Applied Computing. Fortaleza, Brazil. 2008. pp. 1132-1136.

[MR07] Mizzaro, S.; Robertson, S.: HITS hits TREC – exploring IR evaluation results with network analysis. In: Proc. 30th Annual Intl. ACM Conf. on Research and Development in Information Retrieval (SIGIR) Amsterdam. 2007. 479-486

[MWZ07] Moffat, A.; Webber, W.; Zobel, J. Strategic System Comparisons via Targeted Relevance Judgments. In Proc. 30th Annual Intl. ACM Conf. on Research and Development in Information Retrieval (SIGIR) Amsterdam. 2007. 375-382.

[Nu07] Di Nunzio, G.; Ferro, N.; Mandl, T.; Peters, C. CLEF 2006: Ad Hoc Track Overview. In: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, Revised Selected Papers. Berlin et al.: Springer [LNCS 4730] 2007. 21-34.

[Ro06] Robertson, S.: On GMAP: and other transformations. Proc 15th ACM Intl. Conf. on Information and Knowledge Management (CIKM) Arlington, VA. 2006. pp. 872-877.

[SZ05] Sanderson, M.; Zobel, J. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In Proc 28th Annual Intl. Conf. on Research and Development in Information Retrieval (SIGIR) Salvador, Brazil. 2005. 162-169.

[Vo05] Voorhees, E. The TREC robust retrieval track. ACM SIGIR Forum 39 (1) 2005. 11-20.