

Predicting the Effectiveness of Queries and Retrieval Systems

CLAUDIA HAUFF

Chapter 5

A Case for Automatic System Evaluation

5.1 Introduction

Ranking retrieval systems according to their retrieval effectiveness *without* relying on costly relevance judgments is a challenging task which was first explored by Soboroff et al. [133]. The motivation for this research stems from the high costs involved in the creation of test collections, coupled with more and larger collections becoming available. As an illustration, while the GOV2 corpus [38], which was introduced to TREC in 2004, contains roughly 25 million documents, the ClueWeb09 corpus, introduced to TREC in 2009, contains already more than one billion documents, a forty-fold increase in size.

Moreover, in a dynamic environment such as the World Wide Web, where the collection [58, 113] and user search behavior change over time [155, 156], regular evaluation of search engines with human relevance assessments is not feasible [79, 133]. If it becomes possible to determine the relative effectiveness of a set of retrieval systems, reliably and accurately, without the need for relevance judgments, then the cost of evaluation could be greatly reduced.

Additionally, such estimated ranking of retrieval systems could not only serve as a way to compare systems, it can also provide useful information for other applications, such as retrieval model selection [159], where we are interested in finding the best retrieval model in a query dependent fashion, or data fusion, where the estimated ranking of systems can be relied upon to derive merging weights for each system [160, 167].

In recent years, a number of *system ranking estimation* approaches have been proposed [9, 114, 133, 135, 161], which attempt to rank a set of retrieval systems (for a given topic set and test corpus) without human relevance judgments. All these approaches estimate a performance based ranking of retrieval systems by considering the relationship of the top retrieved documents across all or a number of selected systems. While the initial results highlighted the promise of this new direction, the utility of system ranking estimation methods remains unclear, since they have been shown to consistently underestimate the performance of the best systems,

an observation which is attributed to the “tyranny of the masses” effect [9]. This is a very important limitation, as, in practice, it is often the best systems that we are most interested in identifying accurately, rather than the average systems.

In the analysis presented in this chapter, we will show that the problem of mis-ranking the best systems is not inherent to system ranking estimation methods. In previous work [9, 114, 133, 135, 161], the evaluations were mostly performed on the TREC- $\{3,5,6,7,8\}$ data sets. Note that when we refer to a TREC data set, such as TREC-3, we mean all retrieval runs submitted to TREC for the topics of that task. Since a retrieval run is the output of a retrieval system, by ranking retrieval runs, we rank retrieval systems. We will use the terms *run* and *system* mostly interchangeably. In our work, we evaluate system ranking estimation methods on a much wider variety of data sets than previously. We consider a total of sixteen different TREC data sets. They include a range of non-adhoc task data sets, such as expert search [42] and named page finding [43], as well as adhoc tasks on non-traditional corpora, for instance the Blog [115] and Genomics [75] corpora. We observe that the extent of mis-ranking the best systems varies considerably between data sets and is indeed strongly related to the degree of human intervention in the best runs of a data set. This finding suggests that under certain conditions, automatic system evaluation is a viable alternative to human relevance judgments based evaluations.

In a second set of experiments, we also investigate the number of topics required to perform system ranking estimation. In all existing approaches, the retrieval results of the full TREC topic set are relied upon to form an estimate of system performance. However, in [61, 109] it is demonstrated that some topics are better suited than others to differentiate the performance of retrieval systems. Whilst these works were not performed in the context of system ranking estimation, we consider this observation as a starting point for our work. We hypothesize, that with the right subset of topics, the current methods for estimating system rankings without relevance judgment can be significantly improved. To verify this claim, we implement five different approaches [50, 114, 133, 135] to system ranking estimation and compare their performances to each other. We experimentally evaluate the extent of the topic dependent performance and perform a range of experiments to determine the degree to which topic subsets can improve the performance of system ranking estimation approaches. Finally, we attempt to automatically find a good subset of topics to use for system ranking estimation.

Specifically, in this chapter we will show that:

- the ability to accurately identify the best system of a set of retrieval systems is strongly related to the amount of human intervention applied to the system: the larger the amount of human intervention, the less able we are to identify it correctly,
- across the evaluated system ranking estimation approaches, the original work by Soboroff et al. [133] is the most consistent and gives the best performance overall,
- the ability of system ranking estimation methods to estimate a ranking of systems *per topic* varies highly,

- topic subset selection improves on average the performance of the approach proposed by Soboroff et al. [133] by 26% and up to a maximum of 56% (similar improvements are observed for the other system ranking estimation methods), and,
- on average, a subset size of 10 topics yields the highest system ranking estimation performance, a result that is consistent across all data sets and corpora, independent of the number of topics contained in the full TREC topic set.

This chapter is organized as follows: in Section 5.2, we provide an overview of related work in the area of system ranking estimation. Then, in Section 5.3, we introduce the motivation for our experiments in topic subset selection. In Section 5.4, the experimental setup is described and the data sets under investigation are outlined. The empirical analysis, which forms the main part of this chapter, is described in Section 5.5. It contains a comparison of different ranking estimation approaches on the full set of topics (Section 5.5.1) and an analysis of the methods' abilities to determine the correct ranking of systems for each individual topic (Section 5.5.2). The amount of possible performance gain when relying on a subset of topics is discussed in Section 5.5.3. A first attempt to automatically find a good subset of topics is then made in Section 5.5.4. The chapter concludes with a summary in Section 5.6.

5.2 Related Work

Research aiming to reduce the cost of evaluation has been conducted along two lines. Specifically, a number of approaches focus on *reducing* the amount of manual assessments required, while others rely on *fully automatic* evaluation, foregoing the need for manual assessments altogether. Approaches in the first category include the determination of good documents to judge [33, 153], the proposal of alternative pooling methods [8] in contrast to TREC's depth pooling, the proposal of alternative evaluation measures for incomplete judgments [8, 27], the usage of term relevance judgments instead of document relevance judgments [6] and the reliance on manually created queries to derive pseudo-relevant documents [55].

Whilst the aforementioned methods have been developed in the context of TREC, the works by Abdur Chowdhury and his colleagues [16, 17, 37, 79] investigate the evaluation of Web search engines through automatically generated known item queries from query logs and manually built Web directories such as the ODP. Despite the fact, that the queries are derived automatically, we still consider these approaches to be part of the efforts aimed at reducing the amount of manual assessments, as the ODP is constantly maintained by human editors.

In this chapter, we consider approaches of the second category, that is, we focus on algorithms that are completely automatic and require no manual assessments at all. The first work in the ranking of retrieval systems which did not include manually derived relevance judgments is attributed to Soboroff et al. [133] and was motivated by the fact that the relative ranking of retrieval systems remains largely

unaffected by assessor disagreement in the creation of relevance judgments [149]. This observation led to the proposed use of automatically created *pseudo* relevance judgments. In this case, the pseudo relevant documents are derived in the following manner: first, the top retrieved documents across the TREC runs for a particular topic are pooled together such that a document that is retrieved by x systems, appears x times in the pool¹. Then, a number of documents are drawn at random from the pool; those are now considered to be the relevant documents. This process is performed for each topic and the subsequent evaluation of each system is performed with pseudo relevance judgments instead of relevance judgments. In the end, a system's effectiveness is estimated by its pseudo mean average precision. To determine the accuracy of this estimate, the pseudo relevance judgment based system ranking is compared against the ground truth ranking, that is the ranking of systems according to a retrieval effectiveness measure such as MAP. All experiments reported in [133] were performed on the data sets TREC- $\{3,5,6,7,8\}$. Although the reported correlations are significant, one major drawback was discovered: whereas the ranking of the poorly and moderately performing systems is estimated quite accurately with this approach, the ranks of the best performing systems are always underestimated. This is not a small issue, the extent of mis-ranking the best system(s) is severe. For instance, in the TREC-8 data set, where 129 systems are to be ranked, the best system according to the ground truth in MAP is estimated to be ranked at position 113 (Table 5.2), which means it is estimated to be one of the worst performing systems. It was later suggested by Aslam and Savell [9] that this observation can be explained by the “tyranny of the masses” effect, where the best systems are estimated to perform poorly due to them being different from the average. The evaluation can therefore be considered to be based more on popularity than on performance.

The exploitation of pseudo relevant documents has been further investigated by Nuray and Can [114], on a very similar data set, specifically TREC- $\{3,5,6,7\}$. In contrast to Soboroff et al. [133], not all available retrieval systems participate in the derivation of pseudo relevance judgments. The authors experiment with different approaches to find a good subset of $P\%$ of systems. Overall, the best approach is to select those systems that are most different from the average system. Once a subset of systems is determined, the top b retrieved documents of each selected system are merged and the top $s\%$ of the merged result list constitute the pseudo relevance judgments. Different techniques are evaluated for merging the result lists; the best performing approach is to rely on Condorcet voting where each document in the list is assigned a value according to its rank. In this way, it is not only the frequency of occurrence of a document in various result lists that is a factor as in [133], but also the rank the document is retrieved at. By placing emphasis on those systems, that are most dissimilar from the average, this work directly addresses the “tyranny of the masses” criticism [9] levelled at the approach in [133]. The results reported in [114] outperform the results reported in [133] for the data sets evaluated. However, in this chapter, we perform an extensive evaluation across a

¹Removal of duplicates from the pool was also investigated in [133], but proved to be less successful.

larger number of more recent and varied data sets, and will show that this approach does not always deliver a better performance.

Another direction of research is to directly estimate a ranking of systems based on the document overlap between different result lists, instead of deriving pseudo relevance judgments. Wu and Crestani [161] propose to rank the retrieval systems according to their *reference count*. The reference count of a system and its ranked list for a particular topic is the number of occurrences of documents in the ranked lists of the other retrieval systems; a number of normalized and weighted counting methods are also proposed. Experiments on data sets TREC- $\{3,5,6,7,10\}$ generally yield lower correlations than in [114, 133].

A variation, which relies on the *structure of overlap* of the top retrieved documents between different retrieval systems, is proposed by Spoerri [135]. Spoerri suggests that instead of ranking all systems at once, as done in the previous approaches, it is beneficial to repeatedly rank a set of five randomly chosen systems based on their document overlap structure and average the results across all trials to gain a ranking over all systems. It is hypothesized, that adding all available systems at once creates a considerable amount of “noise” as near duplicate systems are entered, thus boosting some documents in a biased way. While the reported experiments on TREC- $\{3,6,7,8\}$ exhibit considerably higher correlations than previous work, and the best systems are consistently ranked in at least the top half of the ranking, it needs to be emphasized that the results are not directly comparable to earlier work. In [135], the evaluation is based only on automatic short runs with the further restriction that only the best performing run per participating group is included. This means for instance, instead of basing the evaluation of TREC-8 on 129 systems as in [9, 114, 133, 161], only 35 systems are evaluated. We will show, that when including all available systems of a data set in the experiments, this method does not perform better than previously introduced approaches.

Finally, in [34] it is proposed to circumvent the problem of mis-ranking the best systems, by assuming those system to be known and reweighting their contribution towards the pseudo-relevance judgments accordingly. The evaluation, performed on TREC- $\{3,4,5,6,7,8\}$, shows the validity of the approach. Such an approach however leads to a circular argument - we rely on automatic evaluation methods to rank the systems according to their performance, but to do so, we require the best systems to be known in advance.

The aforementioned methods have all assumed that all topics of a TREC topic set are equally useful in estimating the ranking of systems. However, recent research on evaluation which relies on manual judgments to rank systems has found that only a subset of topics is needed [61, 109]. We discuss this in the next section and consider how the same idea can be applied when relevance judgments are not available.

5.3 Topic Subset Selection

In order to explore the relationship between a set of TREC topics and a set of retrieval systems, Mizzaro and Robertson [109] took a network analysis based view.

They proposed the construction of a complete bipartite *Systems-Topic graph* where systems and topics are nodes and a weighted edge between a system and a topic represents the retrieval effectiveness of the pair.

Network analysis can then be performed on the graph, in particular, Mizzaro & Robertson employed HITS [87], a method that returns a hub and authority value for each node. The correspondence between those measures and systems and topics was found to be as follows

- the authority of a system node indicates the system’s effectiveness,
- the hubness of a system node indicates the system’s ability to estimate topic difficulty,
- the authority of a topic node is an indicator for topic difficulty, and, finally,
- the hubness of a topic node indicates the topic’s ability to estimate system performance.

While the study in [109] was more theoretic in nature, a recent follow up on this work by Guiver et al. [61] has shown experimentally that when selecting the right subset of topics, the resulting relative system performance is very similar to the system performance on the full topic set, thus allowing the reduction of the number of topics required. The same work though concedes concrete ideas of how to select those topics to future work. Mizzaro [108] also proposed a novel evaluation metric, the Normalized MAP value, which takes the difficulty of a topic into account when evaluating systems.

The finding that individual topics vary in their ability to indicate system performance provides the basis for our work as it implies that there might be subsets of topics that are as suited to estimate the system performance as the full set of topics provided for a TREC task. While the motivation in [61, 109] is to reduce the cost of evaluation by reducing the topic set size, in this work, we are motivated by the fact that system ranking estimation does not perform equally well across all topics.

We examine the following research questions:

- By reducing the topic set, can the performance of current system ranking estimation methods be improved and if so to what extent?
- Can the reduced topic sets that improve the estimation be selected automatically?
- To what extent does the performance of system ranking estimation approaches depend on the set of systems to rank and the set of topics available?

5.4 Materials and Methods

In order to improve the validation power of our results we conduct our analysis on sixteen data sets and five system ranking estimation methods. We shall refer to the estimation methods that we employ in the following fashion:

- the *Data Fusion (DF)* approach by Nuray and Can [114],
- the *Random Sampling (RS)* approach by Soboroff et al. [133],
- the *Structure of Overlap (SO)* approach by Spoerri [135],
- the *Autocorrelation based on Document Scores (ACScore)* approach by Diaz [50], and,
- the *Autocorrelation based on Document Similarity and Scores (ACSimScore)* approach by Diaz [50].

Whereas the first three approaches have already been introduced in Section 5.2, the latter two (*ACSimScore* and *ACScore*) have not been applied to system ranking estimation yet. They are proposed in [50] to evaluate aspect EA3 of Figure 1.1. The main motivation for evaluating specifically those approaches is their mix of information sources. In particular, *RS* relies on document overlap as shown in [9], *SO* considers small subsets of systems for ranking, while *ACScore* and *DF* take the particular retrieval score and the rank respectively a system assigns to a document into account. Finally, the *ACSimScore* approach goes a step further and considers the content similarity between ranked documents and the retrieval scores to determine the relative effectiveness of a system.

Each approach derives a performance score for each pair (t_i, s_j) of topic t_i and system s_j . In order to derive a system's performance score over a particular topic set, the scores the system achieves across all topics in the set are averaged. Based on the scores, assigned to each system by a system ranking estimation approach, the ranking of retrieval systems is estimated. This ranking is then correlated against the ground truth ranking of systems. In our experiments, we will rely on two ground truth rankings of systems:

- Foremost, we rely on the ground truth ranking based on the retrieval effectiveness measure over the entire topic set, which corresponds to aspect EA4 of Figure 1.1. In most instances, this is the ranking of systems according to MAP. Estimating this ranking correctly is the ultimate goal of system ranking estimation approaches. It is utilized in most experiments, the only exception being the experiments in Section 5.5.2.
- In Section 5.5.2, we are interested in how well the ranking of systems can be estimated for each individual topic. Thus, the ground truth ranking is based on the retrieval effectiveness measure of a single topic, which corresponds to aspect EA3 of Figure 1.1. In the experiments of that section, for all but two data sets, the ground truth is the ranking of systems according to average precision. This ranking may or may not coincide with the system ranking based on the retrieval effectiveness measure over the full topic set.

Since we are interested in the ranking of retrieval systems, the evaluation is performed by reporting the rank correlation coefficient Kendall's τ .

5.4.1 Data Sets

As previous work, we also rely on TREC adhoc tasks over different years in our experiments. However, whereas earlier studies focused mainly on TREC- $\{3,5,6,7,8\}$, we investigate a wider variety of data sets, that include more recent adhoc task data sets, a range of non-adhoc task data sets, as well as adhoc tasks on non-traditional corpora. In the previous chapters, we restricted our experiments to three corpora, namely TREC Vol. 4+5, WT10g and GOV2 and their corresponding adhoc retrieval tasks. The main reasons are the availability of the corpora and the importance of the adhoc retrieval task. In the context of system ranking estimation, however, corpus information or training data is not always required. This is the case for the *RS* approach for instance, which relies exclusively on the document identifiers of the top retrieved documents to determine a ranking of retrieval systems. Such a document-content independent approach makes it possible to include a larger number of data sets. In the experiments in this chapter, we take advantage of this fact and evaluate a cross-section of different tasks that have been introduced to TREC over the years. All data sets, that is all the runs submitted by the groups participating in TREC, can be downloaded from the TREC website.

In particular, all experiments are performed on the following TREC data sets:

- **TREC- $\{6,7,8\}$** : adhoc retrieval tasks on TREC Vol. 4+5 [148],
- **TREC- $\{9,10\}$** : adhoc retrieval tasks on WT10g [132],
- **TB- $\{04,05,06\}$** : adhoc retrieval tasks on the TeraByte corpus GOV2 [38],
- **CLIR-01**: the Cross Language track of 2001 [60] which aims at retrieving Arabic documents from a mix of English, French and Arabic topics,
- **NP-02**: the Named Page finding task of 2002 [43] where the task is to find a particular page in a corpus of Web documents,
- **EDISC-05**: the Enterprise Discussion search task of 2005 [42] which relies on a test corpus of e-mails and aims to retrieve e-mails that discuss positive and negative aspects of a topic,
- **EEXP-05**: the Enterprise Expert search task of 2005 [42] which focuses on finding people who are experts on a topic area,
- **BLTR-06**: the Blog track [115], introduced in 2006 to TREC with its topic relevance task as an adhoc-style task on a corpus of blog entries,
- **GEN-07**: the Genomics track of 2007 [75] which focuses on entity based question answering tasks on a collection of biomedical journal articles²,
- **LEGAL-07**: the Legal track of 2007 [141], a recall-oriented track which centers around searching documents in regulatory and litigation settings, and,
- **RELFB-08**: the Relevance Feedback track [26] which intends to study the effects of relevance feedback when different amounts of true relevance feedback is available.

²Although the Genomics task itself calls for passage retrieval, the submitted runs were also evaluated on the document level. The latter interpretation is the one we rely on in our evaluation.

As in Chapter 4 we use a different notation and terminology from Chapters 2 and 3 (data set instead of topic set and TREC-6 instead of 301-350 for instance) to distinguish the current experiments on system ranking estimation from the earlier experiments on query effectiveness prediction.

In all but two data sets, the retrieval effectiveness of a system is measured in MAP. In the NP-02 data set, where the ranking of one particular document is of importance, mean reciprocal rank is the evaluation measure of choice, while in the RELFB-08 data set, the effectiveness measure is statistical MAP [8].

The number of retrieval systems to rank varies between a minimum of 37 (EEXP-05) and a maximum of 129 (TREC-8). The number of topics in the topic set ranges from 25 topics for CLIR-01 to 208 topics for RELFB-08. A comprehensive overview of the number of topics and systems for each data set can be found in Table 5.1. We include all available runs in our experiments, automatic as well as manual and short as well as long runs. This also includes runs, that are not part of the official TREC runs, but are nevertheless available from the TREC website. In the setting of TREC, a run is labelled automatic, if no human intervention was involved in its creation, otherwise it is considered to be manual. The amount of human intervention in manual runs varies, it can range from providing explicit relevance judgments and manually re-ranking documents to small changes in the topic statement to make it accessible for a specific retrieval system. Runs are also categorized as short or long according to the TREC topic part they employ, either the title, description or narrative.

Only one of the approaches we evaluate, namely *ACSimScore*, is based on document content. As in earlier chapters, we preprocessed the corpora by applying Krovetz stemming [90] and stopword removal.

5.4.2 Algorithms

The following sections introduce six system ranking estimation approaches, four of which - *DF*, *RS*, *SO* and *ACScore* - will be investigated throughout this chapter. The document-content based approach *ACSimScore* is only used for comparison in the first set of experiments, as we only have the corpora available for TREC-{6-10} and TB-{04-06}. The system similarity approach by Aslam and Savell [9] is also briefly described, but excluded from further analysis as it is very similar in spirit to *RS*. It also relies on document overlap without consideration the rank or retrieval score of a document, while at the same time being less effective than *RS*.

Data Fusion (*DF*)

We implemented the variation of the data fusion approach, that performed best in [114], namely Condorcet voting and biased system selection. In this approach, a number of parameters need to be set, specifically, (i) the percentage $P\%$ of systems to select in a biased way that favors non-average systems for data fusion, (ii) the number b of top retrieved documents to select from each selected system and (iii) the percentage $s\%$ of documents to use as pseudo relevance judgments from the

merged list of results. We evaluated a range of values for each parameter:

$$\begin{aligned} s &= \{1\%, 5\%, 10\%, 20\%, 30\%, 40\%, 50\%\} \\ b &= \{10, 20, 30, \dots, 100, 125, 150, 175, 200, 250\} \\ P &= \{10\%, 20\%, 30\%, \dots, 100\%\}. \end{aligned}$$

Each of the 1050 possible parameter combinations was tested. To determine the best parameter settings for each data set, we trained on the remaining data sets available for that corpus, that means for instance, that the parameters of the TREC-6 data set were those that led to the best ranking estimation performance for data sets TREC-7 and TREC-8. Depending on the training data sets, widely different parameter combinations were learned, for instance TREC-9 is evaluated on $s = 1\%$, $b = 250$ and $P = 50\%$ while TB-05 is run with $s = 50\%$, $b = 60$ and $P = 80\%$.

Data sets for training are only available for TREC- $\{6-10\}$ and TB- $\{04-06\}$ though. For the remaining data sets, we evaluated all combinations of parameter settings that were learned for TREC- $\{6-10\}$ and TB- $\{04-06\}$. We evaluated each setting on the eight data sets without training data (CLIR-01 to RELFB-08) and chose the setting that across these data sets gave the best performance: $s = 10\%$, $b = 50$ and $P = 100\%$. Thus, the best results are achieved when *not* biasing the selection of systems ($P = 100\%$) towards non-average systems. Since in effect, we optimized the parameters on the test sets, we expect *DF* to perform very well on those data sets.

Random Sampling (*RS*)

We follow the methodology from [133] and rely on the 100 top retrieved documents per retrieval system. We pool the results of *all* systems that are to be ranked, not just the official TREC runs. The percentage of documents to sample from the pool is sampled from a normal distribution with a mean according to the mean percentage of relevant documents in the relevance judgments and a standard deviation corresponding to the deviation between the different topics. Note, that this requires some knowledge about the distribution of relevance judgments; this proved not to be problematic however, as fixing the percentage to a small value (5% of the number of unique documents in the pool) actually yielded little variation in the results. As in [133], due to the inherent randomness of the process, we perform 50 trials. In the end, we average the pseudo average precision values for each pair (t_i, s_j) of topic and system and rank the systems according to pseudo mean average precision.

System Similarity

A simplification of the *RS* process was proposed by Aslam and Savell [9], who observed that retaining duplicate documents in the pool leads to a system ranking estimate that is geared towards document popularity. That is, systems that retrieve many popular documents, are assigned top ranks. Put differently, the more similar a system is to all other systems, the more popular documents it retrieves and thus the better its rank is estimated to be.

The similarity between two systems s_i and s_j is determined by the document overlap of their respective ranked lists of documents R_i and R_j , expressed by the Jaccard similarity coefficient:

$$\text{SysSimilarity}(s_i, s_j) = \frac{|R_i \cap R_j|}{|R_i \cup R_j|}.$$

The estimated effectiveness score of a system s_o is then the average over all pairwise similarities:

$$\text{Avg}(s_o) = \frac{1}{n-1} \sum_{s_i \neq s_o} \text{SysSimilarity}(s_o, s_i).$$

A system's estimated score decreases with decreasing similarity towards the average system.

Structure of Overlap (SO)

Recall, that the structure of overlap approach [135], in contrast to the previously introduced approaches, does not rank all systems at once, but instead repeatedly ranks random sets of five systems. Let there be n systems to be ranked. A total of n random groupings of five systems each are then created such that each system appears in exactly five groupings. Subsequently, for each grouping and for each of the topics, the percentage %Single of documents in the ranked lists of the top retrieved 50 documents found by only one and the percentage %AllFive of documents found by all five systems is determined. The three scores of %Single, %AllFive and the difference (%Single – %AllFive) were proposed as estimated system score. These scores are further averaged across all topics. Since each system participates in five groupings, the scores across those groupings are again averaged, which leads to the final system score.

Autocorrelation based on Document Scores (ACScore)

This approach, proposed by Diaz [50] and denoted by $\rho(\mathbf{y}, \mathbf{y}_\mu)$ in his work, is based on document overlap and the particular retrieval scores a system assigns to each document. Essentially, a retrieval system is estimated to perform well if its scores are close to the average scores across all systems. First of all, the document scores are normalized in order to make them comparable across systems [110]. Then, the average system vector of scores \mathbf{y}_μ is determined as follows: a set \mathcal{U} of the top 75 retrieved documents of all systems is formed and the average score for each element in the set is calculated. Thus the length of vector \mathbf{y}_μ is $m = |\mathcal{U}|$. The linear correlation coefficient r between \mathbf{y}_μ and the vector \mathbf{y} of scores of the top m retrieved documents per system is then the indicator of a system's estimated quality where r is high when both vectors are very similar to each other.

Autocorrelation based on Document Similarity and Scores (*ACSimScore*)

We also evaluate a second approach by Diaz [50] which combines the *ACSim* approach (introduced in Chapter 3, Section 3.2.2) with *ACScore*. This method, originally referred to as $\rho(\tilde{\mathbf{y}}, \mathbf{y}_\mu)$, is based on the notion that well performing systems are likely to fulfill the cluster hypothesis, while poorly performing systems are not. Based on a document's score vector \mathbf{y} , a perturbed score vector $\tilde{\mathbf{y}}$ is derived, which is based on the similarity between the ranked documents of a system. Each element y_i is replaced by the weighted average of scores of the 5 most similar documents (based on TF.IDF) in the ranked list. If the cluster hypothesis is fulfilled, we expect that the most similar documents will also receive a similar score from the retrieval system (y_i and \tilde{y}_i will be similar), while in the opposite case, high document similarity is not expressed in similar scores and \tilde{y}_i will be different from y_i . To score each system, the linear correlation coefficient between \mathbf{y}_μ and the average system vector of scores, $\tilde{\mathbf{y}}$, is determined.

5.5 Experiments

In Section 5.5.1, we compare the introduced ranking estimation approaches across the different data sets. Then, in Section 5.5.2, we will show that the system ranking cannot be estimated equally well for each topic. In Section 5.5.3, we perform a number of motivational experiments to determine whether it is possible to exploit this observation. Finally, in Section 5.5.4, we make a first attempt at automatically selecting a good subset of topics from the full topic set.

5.5.1 System Ranking Estimation on the Full Set of Topics

In this section, we replicate the experiments reported in [114, 133]. In contrast to [135], we apply *SO* to rank all available systems per data set. Additionally, we investigate *ACSimScore* and *ACScore* for their ability to rank retrieval systems, instead of ranking systems for single topics as reported in [50]. The results of our experiments are shown in Table 5.1. The highest correlation achieved for each data set is given in bold; the correlations that are not significantly different from the best one are underlined. All correlations reported are significantly different from zero with a p-value < 0.005 .

When comparing the correlations in Table 5.1 with those reported in earlier chapters for the task of predicting the query effectiveness for a particular retrieval system (Figure 1.1, EA2), it is evident that the task of estimating a ranking of retrieval systems is less difficult to achieve. The weakest approach, *ACSimScore* estimates the ranking of systems with a correlation between $\tau = 0.42$ and $\tau = 0.65$. Noteworthy are the high correlations the five estimators achieve on the TREC- $\{9,10\}$ data sets in comparison to the TREC- $\{6,7,8\}$ data sets. For the task of query effectiveness estimation the opposite observation was made in previous chapters: the performance of queries of TREC Vol. 4+5 can be predicted very well, while

| | #sys | #top | Kendall's Tau | | | | |
|-----------------|------|------|---------------|--------------|--------------|--------------|--------------|
| | | | DF | ACSimScore | ACScore | SO | RS |
| TREC-6 | 73 | 50 | 0.600 | 0.425 | 0.429 | 0.470 | 0.443 |
| TREC-7 | 103 | 50 | 0.486 | <u>0.417</u> | <u>0.421</u> | <u>0.463</u> | <u>0.466</u> |
| TREC-8 | 129 | 50 | 0.395 | 0.467 | 0.438 | <u>0.532</u> | 0.538 |
| TREC-9 | 105 | 50 | 0.527 | <u>0.639</u> | <u>0.655</u> | 0.634 | 0.677 |
| TREC-10 | 97 | 50 | <u>0.621</u> | <u>0.649</u> | 0.663 | 0.598 | <u>0.643</u> |
| TB-04 | 70 | 50 | 0.584 | 0.647 | <u>0.687</u> | 0.614 | 0.708 |
| TB-05 | 58 | 50 | <u>0.606</u> | 0.574 | 0.547 | <u>0.604</u> | 0.659 |
| TB-06 | 80 | 50 | <u>0.513</u> | <u>0.458</u> | 0.528 | <u>0.447</u> | <u>0.518</u> |
| CLIR-01 | 47 | 25 | <u>0.697</u> | - | <u>0.700</u> | <u>0.650</u> | 0.702 |
| NP-02 | 70 | 150 | <u>0.667</u> | - | 0.696 | <u>0.668</u> | <u>0.693</u> |
| EDISC-05 | 57 | 59 | 0.668 | - | 0.560 | <u>0.614</u> | <u>0.666</u> |
| EEXP-05 | 37 | 50 | <u>0.589</u> | - | 0.682 | 0.502 | 0.483 |
| BLTR-06 | 56 | 50 | <u>0.482</u> | - | <u>0.485</u> | 0.357 | 0.523 |
| GEN-07 | 66 | 36 | 0.578 | - | 0.500 | 0.362 | <u>0.563</u> |
| LEGAL-07 | 68 | 43 | 0.754 | - | 0.680 | <u>0.749</u> | <u>0.741</u> |
| RELFB-08 | 117 | 208 | 0.537 | - | 0.599 | 0.544 | 0.559 |

Table 5.1: System ranking estimation on the full set of topics. Reported is Kendall's τ . All correlations reported are significant ($p < 0.005$). The highest correlation per topic set is bold. The correlations that are not statistically different from the best one are underlined. Column #sys shows the number of systems to rank, #top shows the number of topics in a data set.

for the queries of WT10g predicting the effectiveness is difficult. As will be shown later in this chapter, system ranking estimation is more difficult on TREC- $\{6,7,8\}$ due to the greater amount of human intervention in the best runs. Manual runs can be very different from automatic runs, containing many unique documents that are not retrieved by other systems. This is a problem as system ranking estimation is to some extent always based on document popularity. A simple solution would be to prefer runs, that retrieve many unique documents, however this is not possible since the worst performing runs also retrieve a lot of documents that are not retrieved by any other run.

DF outperforms RS on TREC- $\{6,7\}$ as already reported in [114]. The poor result on TREC-8 is due to an extreme parameter setting found to perform best on TREC- $\{6,7\}$, which was subsequently used to evaluate DF on TREC-8. On the remaining data sets where DF 's parameters were trained (TREC- $\{9,10\}$ and TB- $\{04,05,06\}$), RS outperforms DF , in two instances significantly. The highly collection dependent behavior of DF is due to the method's inherent bias in the way in which the subset of systems to select the pseudo relevant documents from are determined. A system that is dissimilar to the average system, can either perform very well or very poorly. On the data sets without training data (CLIR-01 to RELFB-08), DF performs similarly to RS , which is not surprising as the best performing parameter setting of DF means that the most popular documents are included in

the pseudo relevant documents, just as for *RS*. The only exception is data set EEXP-05, where *DF* achieves a correlation of $\tau = 0.59$, while *RS* achieves a correlation of $\tau = 0.48$. This variation can be explained by the small number of systems to rank (37) - here a small difference in system rankings has a considerable effect on Kendall's τ .

Relying on TFIDF based content similarity does not help, shown by *ACSimScore*'s performance. In four out of eight evaluated data sets, its performance is significantly worse than the best performing approach. Although the approach ranks systems higher that are closer to the average (just like *RS*), the TFIDF similarity might not be reflected in the score similarity as is expected in this approach. In particular more advanced retrieval systems are likely to include more than basic term statistics such as evidence from external corpora, anchor text, the hyperlink structure, etcetera, all of which influences the retrieval scores a system assigns to a document.

The *SO* approach performs similarly to *RS* on TREC- $\{6,7,8,9\}$, while for the remaining data sets the differences in performance generally become larger. We suspect that the ranking of five systems is less stable, than the ranking of all systems at once.

ACScore, which takes the score a retrieval system assigns to a document into account, is also well performing, for five data sets it achieves the highest correlation. A potential disadvantage of *ACScore* though is that it requires knowledge of the retrieval scores a retrieval system assigns to each document, while *DF*, *SO* and *RS* require no such knowledge.

In Table 5.1 we have refrained from reporting a mean correlation across all data sets for each estimator on purpose, due to the different data set sizes, that is the number of retrieval systems to rank. Instead, we point out, that *RS* exhibits the highest correlation for six data sets, while *DF* and *ACScore* record the highest correlation on five data sets each. Additionally, *RS*'s performance is significantly worse than the best performing approach in only three instances, *DF* is significantly worse in four and *ACScore* is significantly worse in six data sets. Taking into account, that *DF*'s parameters were optimized on eight of the sixteen data sets, we conclude that in contrast to earlier work which was performed on a small number of TREC data sets [9, 114, 135, 161], when evaluating a broader set of data sets, the random sampling approach *RS* is the most consistent and overall the best performing method.

Rank Estimate of the Best Performing System

As discussed in Section 5.2, the commonly cited problem of automatic system evaluation is the mis-ranking of the best systems. As in previous work evaluations have mostly been carried out on TREC- $\{3,5,6,7,8\}$, where the problem of underestimating the best systems occurs consistently, it has been assumed to be a general issue. When considering more recent and diverse data sets, we find this problem to be dependent on the set of systems to rank. To give an impression of the accuracy of the rankings, in Figure 5.1 scatter plots of the estimated system ranks versus the ground truth system ranks are shown for a number of data sets and system ranking

estimation approaches along with the achieved correlation and the estimated rank (ER) of the best system. Each data point stands for one of the n systems and the best system is assigned rank 1 in the ground truth ranking. In the ideal case, when the ranks of all systems are estimated correctly and therefore $\tau = 1.0$, the points would lie on a straight line from $(1, 1)$ to (n, n) .

| | Best Run | M/A | #sys | Estimated Rank | | | |
|----------|---------------------------|-----|------|----------------|---------|-----|-----|
| | | | | DF | ACScore | SO | RS |
| TREC-6 | <i>uwmt6a0</i> [41] | M | 73 | 52 | 55 | 56 | 57 |
| TREC-7 | <i>CLARIT98COMB</i> [56] | M | 103 | 48 | 70 | 78 | 74 |
| TREC-8 | <i>READWARE2</i> [1] | M | 129 | 112 | 117 | 104 | 113 |
| TREC-9 | <i>iit00m</i> [36] | M | 105 | 83 | 79 | 76 | 76 |
| TREC-10 | <i>iit01m</i> [2] | M | 97 | 80 | 84 | 87 | 83 |
| TB-04 | <i>uogTBQEL</i> [120] | A | 70 | 23 | 26 | 30 | 30 |
| TB-05 | <i>indri05Admfl</i> [106] | A | 58 | 35 | 45 | 30 | 32 |
| TB-06 | <i>indri06AtdnD</i> [105] | A | 80 | 12 | 5 | 28 | 20 |
| CLIR-01 | <i>BBN10XLB</i> [164] | A | 47 | 3 | 2 | 6 | 2 |
| NP-02 | <i>thunp3</i> [173] | A | 70 | 18 | 16 | 20 | 17 |
| EDISC-05 | <i>TITLETRANS</i> [101] | A | 57 | 1 | 2 | 2 | 1 |
| EEXP-05 | <i>THUENTO505</i> [59] | A | 37 | 8 | 3 | 12 | 10 |
| BLTR-06 | <i>wxoqf2</i> [165] | A | 56 | 5 | 4 | 13 | 5 |
| GEN-07 | <i>NLMinter</i> [48] | M | 66 | 1 | 1 | 2 | 1 |
| LEGAL-07 | <i>otL07frw</i> [140] | M | 68 | 4 | 15 | 8 | 4 |
| RELFB-08 | <i>Brown.E1</i> [98] | M | 117 | 64 | 61 | 61 | 65 |

Table 5.2: Estimated rank of the system that performs best according to the ground truth. The evaluation metric of the ground truth is mean reciprocal rank (NP-02), statistical MAP [8] (RELFB-08) and MAP (all other data sets) respectively. M/A indicates if the best system is manual (M) or automatic (A) in nature. #sys shows the number of systems (or runs) to rank. The last four columns depict the estimated rank of the best system. Rank 1 is the top rank.

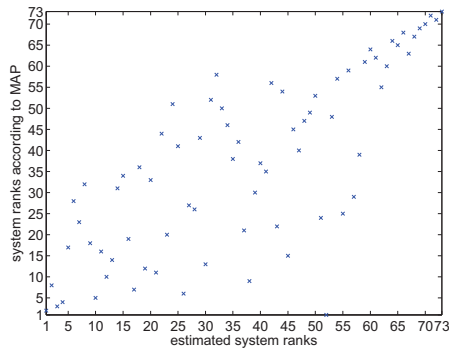
The scatter plots in Figure 5.1a, 5.1b and 5.1c reveal the extent of mis-ranking the best systems of data sets TREC- $\{6,8,10\}$ respectively. In case of TREC-6, only the best system is severely mis-ranked, while in data set TREC-8 the best ten systems are estimated to perform poorly with estimated ranks of 70 or worse, in fact, the best system is estimated to be ranked at position 113 out of 129. In the TREC-10 data set, the two best performing systems are ranked together with the worst performing systems, while the other well performing systems are ranked towards the top of the ranking. When we consider the TB-04 data set (Figure 5.1d), a decrease in the amount of mis-ranking of the best systems is evident. The best correspondence between estimated and ground truth based ranking can be found in Figure 5.1g where the results of data set LEGAL-07 are shown. The correlation of $\tau = 0.75$ indicates the quality of the estimated ranking, and the best system has an estimated rank of four. Better in terms of the best systems performs only *DF* on EDISC-05 (Figure 5.1e), where the two best performing systems are estimated correctly at

ranks one and two.

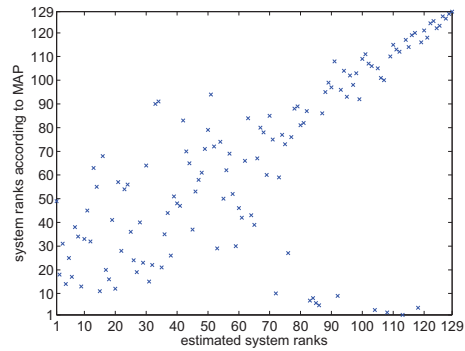
An overview of the estimated rank of the best system across all data sets is given in Table 5.2, where the best rank estimate of a data set is indicated in bold. For comparison purposes, we also list the number of systems to rank once more. Note, that we exclude the approach *ACSimScore* from further experiments, as it is not available for all data sets and moreover has shown to be the weakest performing method on the evaluated data sets. We observe that independent of the system ranking estimation approach, the problem of underestimating the ranking of the best system decreases considerably for the data sets TB-{04-06} in comparison to TREC-{6-10}. With the exception of RELFB-08, the ranks of the best systems are estimated to a much greater accuracy, in fact five for data sets (CLIR-01, EDISC-05, BLTR-06, GEN-07, LEGAL-07) *DF* and *RS* estimate the best system within the top five ranks. When we investigated this discrepancy in estimating the rank of the best system between the different data sets, it became apparent, that the reason for this behavior lies in the makeup of the best run. Table 5.2 also lists the best system of each data set according to the ground truth and an indicator if the best system is manual or automatic in nature. For data sets TREC-{6-10} in all cases, the top performing run according to the MAP based ground truth ranking is manual. The amount of manual intervention in each run is significant:

- In TREC-6, the run *uwmt6a0* [41] was created by letting four human assessors spent a total of 105 hours (2.1 hours on average per topic) on the creation of queries and the judging of documents, which lead to 13064 judgments being made.
- In TREC-7, the run *CLARIT98COMB* [56] was created by having each topic judged by four different assessors. The judged documents were then included as relevance feedback in the final result run, with additional resorting to move the documents manually labeled as relevant to the top of the ranking.
- In TREC-8, the run *READWARE2* [1] was created by letting a retrieval system expert create numerous queries for each TREC topic by considering the top retrieved documents and reformulating the queries accordingly. On average, 12 queries were created per TREC topic.
- In TREC-9, the run *iit00m* [36] was created by letting an expert derive a query with constraints for each TREC topic.
- In TREC-10, the run *iit01m* [2] was created with the aid of manual relevance feedback: the top ranked documents were assessed and reranking was performed accordingly.

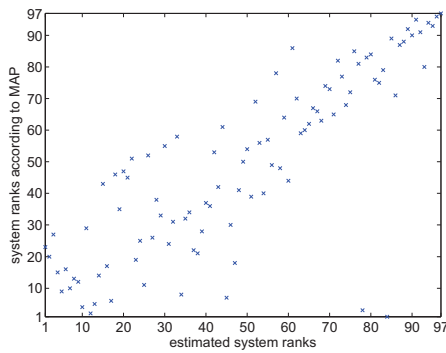
The runs created this way are very different from automatic runs (which one way or another are dependent on the collection frequencies of the query terms) and are bound to have a small amount of overlap in the top retrieved documents in comparison to the automatic runs. This explains why system ranking estimation approaches uniformly estimate them to be among the worst performing runs. In contrast, the estimated ranks of the best systems of GEN-07 and LEGAL-07, which are also classified as manual, are highly accurate. This is explained by the fact,



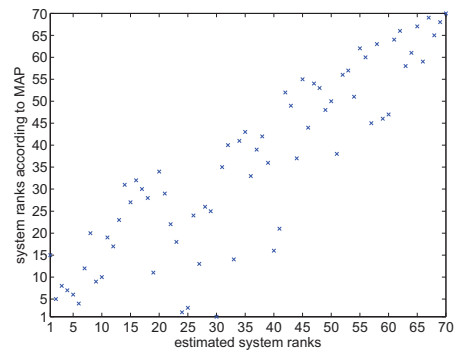
(a) TREC-6, DF , $\tau = 0.60$, $ER = 52$



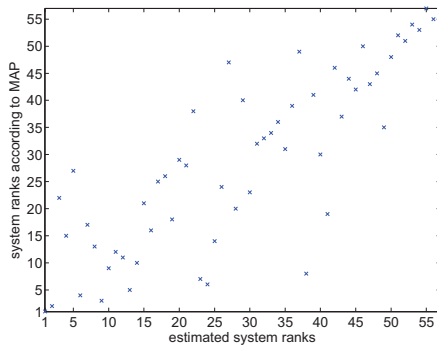
(b) TREC-8, RS , $\tau = 0.54$, $ER = 113$



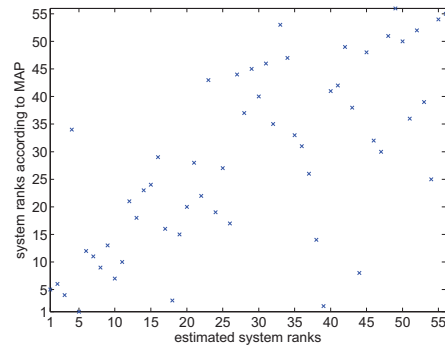
(c) TREC-10, $ACScore$, $\tau = 0.66$,
 $ER = 84$



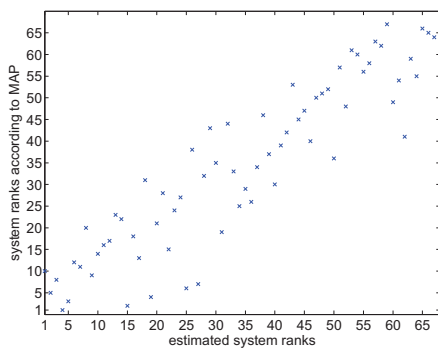
(d) TB-04, RS , $\tau = 0.71$, $ER = 30$



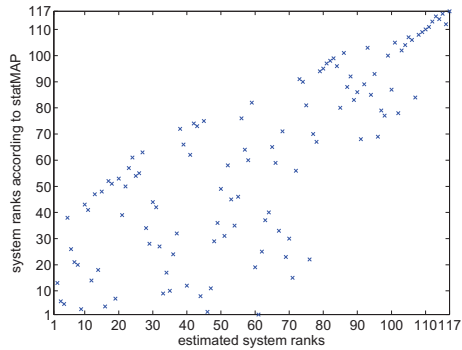
(e) EDISC-05, DF , $\tau = 0.67$, $ER = 1$



(f) BLTR-06, RS , $\tau = 0.52$, $ER = 5$



(g) LEGAL-07, DF , $\tau = 0.75$, $ER = 4$



(h) RELFB-08, $ACScore$, $\tau = 0.60$,
 $ER = 61$

Figure 5.1: Scatter plots of system ranks according to system ranking estimation approaches (x-axis) versus system ranks according to the ground truth (y-axis). Each marker stands for one retrieval system. Rank 1 is assigned to the best system.

that in both instances, the runs were created with little human intervention. The best run of GEN-07, *NLMinter* [48], is tagged as manual, because the Genomics topics were manually transcribed into queries for a domain specific external search engine and the documents retrieved from this engine were used for collection enrichment. The best run of LEGAL-07, *otL07frw* [140], is even less manual. Here, the provided TREC topics were manually transformed into queries suitable for the search engine without adding any additional knowledge by the human query transcriber. Finally, we note that the poor estimation performance on RELFB-08 and its best run, *Brown.E1* [98], is a result of the task [26], where the performance of pseudo-relevance feedback algorithms was investigated, by providing four different sets of relevance judgments, the smallest set with one relevant document per topic, the largest set with between 40-800 judged documents per topic. We hypothesize that based on the very different type of relevance information between the runs, document overlap might not be a good indicator.

A comparison of the performances of DF on data sets TREC-6 and GEN-07 reveals, that reporting both the correlation coefficient τ and the estimated rank of the best performing system offers better insights into the abilities of a system ranking estimation method. For both data sets, DF performs similarly with respect to the rank correlation, $\tau = 0.60$ (TREC-6) and $\tau = 0.58$ (GEN-07) respectively. However, the performances with respect to the estimated rank of the best system are very different, while the estimate of TREC-6 is highly inaccurate ($ER = 52$ out of 73 systems), the best system of GEN-07 is identified correctly ($ER = 1$ out of 66 systems).

Considering the success of estimating the rank of the best system across the four system ranking estimation approaches, we note that DF and $ACScore$ outperform SO and RS by providing the best estimates for seven data sets, while RS and SO provide the best estimates on five and four data sets respectively. In most instances, the estimated ranks are similar across all approaches, exceptions are data sets TREC-7 and TB-06 where the maximum difference in ER is 30 and 23 respectively. Although on TREC-7 all four ranking estimation approaches result in similar correlations (between $\tau = 0.42$ and $\tau = 0.47$), DF 's estimate of the best system is considerably better than of the remaining approaches. This reiterates the previous point, that both τ and ER should be reported to provide a more comprehensive view of an algorithm's performance.

5.5.2 Topic Dependent Ranking Performance

In this section, we show that the ability of system ranking estimation approaches to rank retrieval systems correctly differs significantly between the topics of a topic set. While for a number of topics the estimated rankings are highly accurate and close to the actually observed rankings, for other topics system ranking estimation fails entirely.

We set up the following experiment: for each topic, we evaluated the estimated ranking of systems (Figure 1.1, EA3) by correlating it against the ground truth ranking that is based on average precision, reciprocal rank or statistical average preci-

sion. This is different from the ground truth ranking based on MAP, mean reciprocal rank or statMAP. Here, we are *not* interested in how well a single topic can be used to approximate the ranking of systems over the entire topic set. Instead, we are interested in how well a system ranking estimation method performs for each individual topic. To evaluate the range of performances, we record the topic for which the least correlation is achieved and the topic for which the highest correlation is achieved. The results are shown in Table 5.3.

| | DF | | ACScore | | SO | | RS | |
|-----------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | min. τ | max. τ | min. τ | max. τ | min. τ | max. τ | min. τ | max. τ |
| TREC-6 | 0.008 | 0.849† | -0.161 | 0.812† | -0.147 | 0.752† | -0.105 | 0.814† |
| TREC-7 | -0.061 | 0.765† | 0.053 | 0.695† | -0.008 | 0.764† | -0.004 | 0.693† |
| TREC-8 | 0.053 | 0.792† | 0.087 | 0.740† | 0.080 | 0.723† | 0.143 | 0.731† |
| TREC-9 | -0.234† | 0.835† | 0.018 | 0.760† | 0.096 | 0.760† | 0.179 | 0.730† |
| TREC-10 | -0.094 | 0.688† | 0.031 | 0.722† | 0.054 | 0.707† | 0.130 | 0.821† |
| TB-04 | 0.002 | 0.906† | -0.161 | 0.777† | -0.057 | 0.784† | -0.025 | 0.882† |
| TB-05 | 0.040 | 0.769† | -0.161 | 0.716† | -0.052 | 0.709† | -0.083 | 0.827† |
| TB-06 | -0.070 | 0.728† | 0.055 | 0.644† | -0.159 | 0.710† | -0.152 | 0.760† |
| CLIR-01 | 0.268 | 0.862† | 0.378† | 0.837† | 0.220 | 0.876† | 0.248 | 0.862† |
| NP-02 | -0.264 | 0.607† | -0.129 | 0.760† | -0.239 | 0.621† | -0.257 | 0.649† |
| EDISC-05 | -0.019 | 0.573† | -0.038 | 0.589† | -0.021 | 0.526† | 0.024 | 0.640† |
| EEXP-05 | -0.250 | 0.845† | -0.224 | 0.808† | -0.294 | 0.764† | -0.208 | 0.770† |
| BLTR-06 | 0.044 | 0.534† | 0.018 | 0.507† | -0.192 | 0.436† | 0.206 | 0.562† |
| GEN-07 | 0.151 | 0.795† | 0.040 | 0.700† | 0.078 | 0.627† | 0.180 | 0.774† |
| LEGAL-07 | 0.027 | 0.690† | -0.004 | 0.583† | -0.058 | 0.691† | -0.008 | 0.690† |
| RELFB-08 | -0.183† | 0.797† | -0.115 | 0.749† | -0.172 | 0.774† | -0.137 | 0.775† |

Table 5.3: Topic dependent ranking performance: minimum and maximum estimation ranking accuracy in terms of Kendall’s τ . Significant correlations ($p < 0.005$) are marked with †.

The results are very regular across all data sets and system ranking estimation methods: the spread in correlation between the best and worst case are extremely wide; in the worst case, there is no correlation ($\tau \approx 0$) between the ground truth and the estimated ranking or in rare cases a significant negative correlation is observed (such as for data sets TREC-9 and RELFB-08). In the best case on the other hand, the estimated rankings are highly accurate, and with few exceptions $\tau > 0.7$. Overall, *DF* exhibits the highest correlation on TB-04, where the maximum achievable correlation is $\tau = 0.91$. Though not explicitly shown, we note that the topics for which the minimum and maximum τ are recorded vary between the different system ranking estimation approaches.

These findings form the main motivation for our work: if we were able to determine a subset of topics for which the system ranking estimation algorithms perform well, we hypothesize that this would enable us to achieve a higher estimation accuracy of the true ranking across the full set of topics.

5.5.3 How Good are Subsets of Topics for Ranking Systems?

Having shown in the previous section that the quality of ranking estimation varies across individual topics, we now turn to investigating whether selecting a subset of topics from the full topic set is useful in the context of system ranking estimation algorithms. That is, we attempt to determine whether we can improve the accuracy of the approaches over the results reported in Section 5.5.1 on the full set of topics. We can choose a subset of topics, for instance, by removing those topics from the full set of topics the system ranking approach performs most poorly on. To investigate this point, we experiment with selecting subsets of topics according to different strategies as well as evaluating a large number of randomly drawn subsets of topics.

Each of the evaluated topic sets consists of m topics, m varies from 25 to 208 (Table 5.1). We therefore test subsets of cardinality $c = \{1, 2, \dots, m\}$. In the ideal case, for each cardinality we would test all possible subsets. This is not feasible though, as for each cardinality c , a total of $\binom{m}{c}$ different subsets exist; for a topic set with $m = 50$ topics and subsets of cardinality $c = 6$ for instance this already amounts to nearly sixteen million subset combinations, that is $\binom{50}{6} = 15890700$. For this reason, for each c , we randomly sample 10000 subsets of topics. Apart from this random strategy, we also include a number of iterative topic selection strategies, that will be described shortly.

For the topic subsets of each cardinality, we determine the correlation between the estimated ranking of systems (based on this subset) and the ground truth ranking of systems based on the retrieval effectiveness across the full set of topics. In contrast to Section 5.5.2, we are now indeed interested in how well a subset of one or more topics can be used to approximate the ranking of systems over the entire topic set.

In total, we report results for five subset selection strategies, two based on samples of subsets and three iterative ones:

- **worst sampled subset:** given the 10000 sampled subsets of a particular cardinality c , reported is the τ of the subset resulting in the lowest correlation,
- **average sampled subset:** given the 10000 sampled subsets of a particular cardinality c , reported is the average τ across all samples,
- **greedy approach:** an iterative strategy; at cardinality c , that topic, from the pool of unused topics, is added to the existing subset of $c - 1$ topics, for which the new subset reaches the highest correlation with respect to the ground truth ranking based on MAP; this approach performs usually as well as or better than the best sampled subset, which is therefore not listed separately,
- **median AP:** an iterative strategy; at cardinality c that topic is added to the existing subset of $c - 1$ topics, that exhibits the highest median average precision across all systems; this means that first the easy topics (on which many systems achieve a high average precision) are added and then the difficult ones,
- **estimation accuracy:** an iterative strategy; at cardinality c that topic is added to the existing subset of $c - 1$ topics, that best estimates the ranking of systems

according to average precision for that topic; thus, first those topics are added to the subset that the system ranking estimation method achieves the highest estimation accuracy for (this strategy draws from results of Section 5.5.2).

We should stress here, that the latter three strategies (greedy, median AP and estimation accuracy) all require knowledge of the true relevance judgments. This experiment was set up to determine whether it is at all beneficial to rely on subsets instead of the full topic set. These strategies were not designed to find a subset of topics automatically. Therefore this section should be viewed as an indicator that subset selection is indeed a useful research subject that should be pursued further.

The results of this analysis are shown in Figure 5.2 for selected data sets. After a visual inspection it becomes immediately evident that the general trend of the results is similar across all examples. The greedy approach, especially at small subset sizes between $c = 5$ and $c = 15$, yields significantly higher correlations than the baseline, which is the correlation the method achieves at the full topic set size of m topics. After a peak, the more topics are added to the topic set, the lower the correlation. The amount of change of τ is data set dependent, the largest change in Figure 5.2 is observed for TREC-9 and the *DF* approach, where τ increases from the baseline correlation of $\tau = 0.53$ to $\tau = 0.80$ at the peak of the greedy approach. The worst subset strategy on the other hand shows the potential danger of choosing the wrong subset of topics: τ is significantly lower than the baseline for small cardinalities.

When averaging τ across all sampled subsets (the average subset strategy) of a cardinality, at subset sizes of about $m/3$ topics, the correlation is only slightly worse than the baseline correlation.

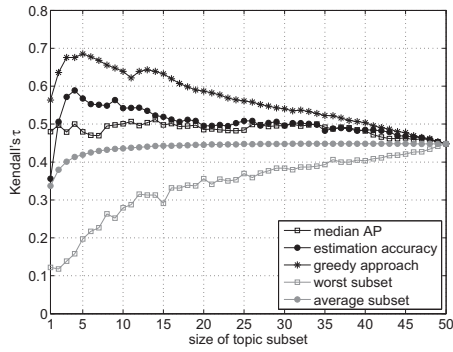
When considering the median AP strategy, which first adds easy topics (topics with a high median AP) to the subset of topics, the gains in correlation over the baseline are visible in Figures 5.2a and 5.2f but they are topic dependent and far less pronounced than the best possible improvement, as exemplified by the greedy approach.

Better than the median AP strategy is the performance of the estimation accuracy strategy, where first those topics are added to the topic subset, for whom the ranking of systems is estimated most accurately as measured by Kendall's τ . This strategy is based on the results of Section 5.5.2. Particularly high correlations are achieved in Figures 5.2b, 5.2c, 5.2e and 5.2f. Here, the development of the correlation coefficient achieved by the estimation accuracy strategy across different cardinalities mirrors the development of the greedy subset approach. However, the improvements are also not consistent across all data sets. In the worst case as seen in Figure 5.2g, the correlations are not better than for the average subset approach. Overall though, the estimation accuracy strategy is also a subset selection method worth pursuing. It remains to be seen though, whether an automatic procedure can be devised that allows us to estimate the accuracy of the ranking to a high degree.

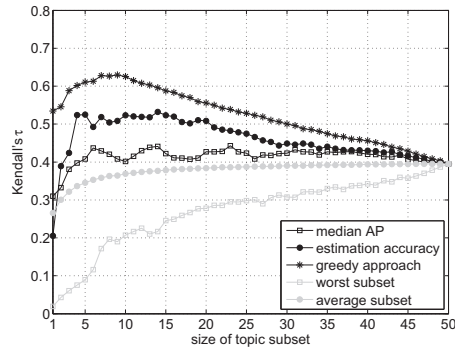
A summary of the most important results of *DF*, *ACScore*, *SO* and *RS* on all data sets are shown in Table 5.4. Listed are the the correlation coefficients on the full set of topics as well as the correlation of the best performing subset in the

| | DF | | ACScore | | SO | | RS | |
|-----------------|----------|---------------------|----------|---------------------|----------|---------------------|----------|---------------------|
| | full set | greedy $\pm\%$ | full set | greedy $\pm\%$ | full set | greedy $\pm\%$ | full set | greedy $\pm\%$ |
| | τ | τ | τ | τ | τ | τ | τ | τ |
| TREC-6 | 0.600 | 0.804 +34.0% | 0.429 | 0.723 +68.5% | 0.470 | 0.731 +55.5% | 0.443 | 0.654 +47.6% |
| TREC-7 | 0.486 | 0.762 +56.8% | 0.421 | 0.591 +40.4% | 0.463 | 0.633 +36.7% | 0.466 | 0.584 +25.3% |
| TREC-8 | 0.395 | 0.630 +59.5% | 0.438 | 0.606 +38.4% | 0.532 | 0.661 +24.2% | 0.538 | 0.648 +20.4% |
| TREC-9 | 0.527 | 0.800 +51.8% | 0.655 | 0.780 +19.1% | 0.634 | 0.775 +22.2% | 0.677 | 0.779 +15.1% |
| TREC-10 | 0.621 | 0.761 +22.5% | 0.663 | 0.755 +13.9% | 0.598 | 0.711 +18.9% | 0.643 | 0.734 +14.2% |
| TB-04 | 0.584 | 0.898 +53.8% | 0.687 | 0.829 +20.7% | 0.614 | 0.804 +30.9% | 0.708 | 0.846 +19.5% |
| TB-05 | 0.606 | 0.800 +32.0% | 0.547 | 0.743 +35.8% | 0.604 | 0.786 +30.1% | 0.659 | 0.812 +23.2% |
| TB-06 | 0.513 | 0.682 +32.9% | 0.528 | 0.707 +33.9% | 0.447 | 0.632 +41.4% | 0.518 | 0.704 +35.9% |
| CLIR-01 | 0.697 | 0.785 +12.6% | 0.700 | 0.815 +16.4% | 0.650 | 0.771 +18.5% | 0.702 | 0.808 +15.0% |
| NP-02 | 0.667 | 0.839 +25.8% | 0.696 | 0.875 +25.7% | 0.668 | 0.838 +25.4% | 0.693 | 0.853 +23.0% |
| EDISC-05 | 0.668 | 0.776 +16.2% | 0.560 | 0.703 +25.5% | 0.614 | 0.773 +25.9% | 0.666 | 0.801 +20.3% |
| FEEXP-05 | 0.589 | 0.900 +52.9% | 0.682 | 0.874 +28.2% | 0.502 | 0.745 +48.5% | 0.483 | 0.718 +48.4% |
| BLTR-06 | 0.482 | 0.617 +28.0% | 0.485 | 0.603 +24.4% | 0.357 | 0.538 +51.0% | 0.523 | 0.601 +14.9% |
| GEN-07 | 0.578 | 0.685 +18.5% | 0.500 | 0.672 +34.5% | 0.362 | 0.569 +57.2% | 0.563 | 0.680 +20.9% |
| LEGAL-07 | 0.754 | 0.864 +14.6% | 0.680 | 0.808 +18.9% | 0.749 | 0.874 +16.6% | 0.741 | 0.865 +16.7% |
| RELFB-08 | 0.537 | 0.878 +63.5% | 0.599 | 0.895 +49.4% | 0.544 | 0.859 +57.9% | 0.559 | 0.872 +56.1% |

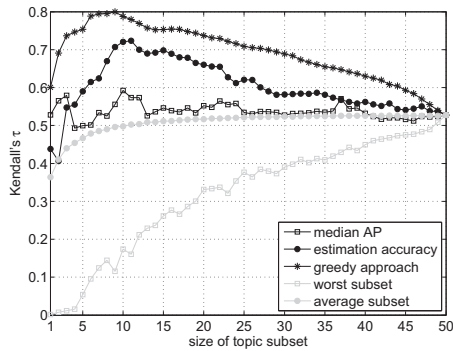
Table 5.4: Summary of topic subset selection experiments. In bold, the highest correlation coefficient of the greedy strategy per topic set. The columns marked with \pm show the percentage of change between τ achieved on the full topic set and the greedy approach. All correlations reported are significant ($p < 0.005$). All differences between the best greedy τ and the τ of the full topic set are statistically significant.



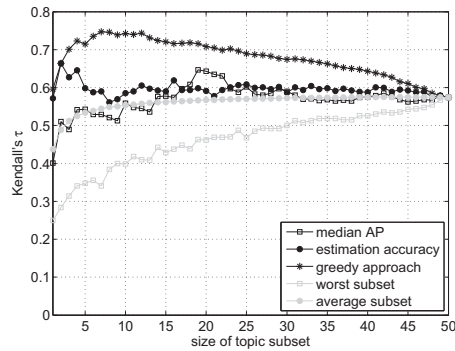
(a) TREC-6, *RS*



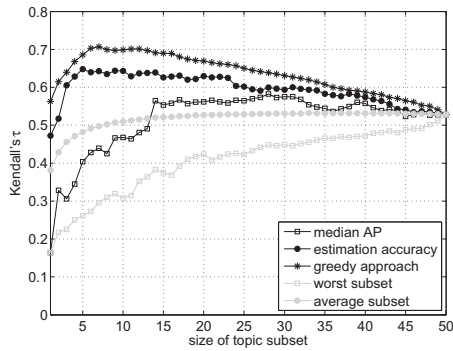
(b) TREC-8, *DF*



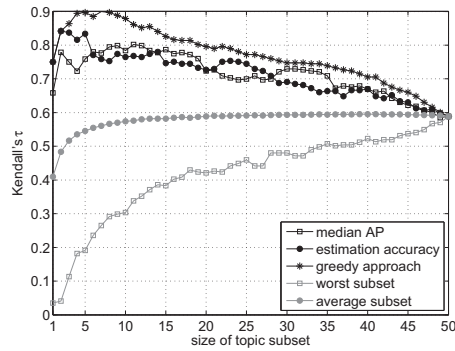
(c) TREC-9, *DF*



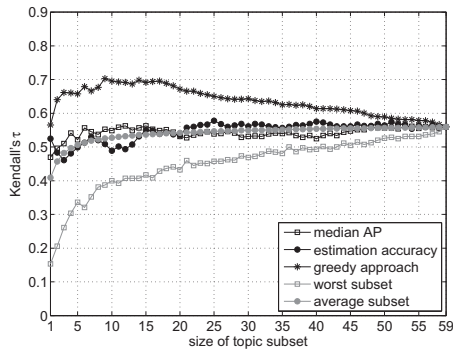
(d) TB-05, *ACSimScore*



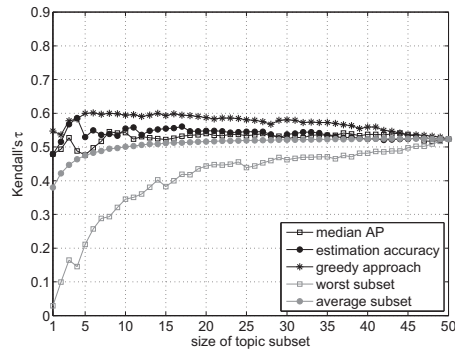
(e) TB-06, *ACScore*



(f) EEXP-05, *DF*



(g) EDISC-05, *ACScore*



(h) BLTR-06, *RS*

Figure 5.2: Topic subset selection experiments.

greedy approach, which is the maximum amount of improvement we assume possible. Though this is not entirely correct, the fact that the best sample among the random subsets does not perform better than the greedy approach suggests, that this is an adequate approximation of the true optimal performance. Across all pairings of system ranking estimation approach and topic set, subsets of topics indeed exist that would greatly improve the performance of system ranking estimation algorithms. Consider for instance, the results of *RS* on RELFB-08: with the “right” topic subset, a rank correlation of $\tau = 0.87$ can be reached, a 56% increase over the performance on the full topic set ($\tau = 0.56$). Given in bold, is the best possible correlation for each data set. Here, the *DF* approach shows the most potential, in particular for eight out of sixteen data sets it records the highest possible improvement.

5.5.4 Automatic Topic Subset Selection

The observations made in the previous two sections can only be useful in practice if it becomes possible to automatically identify those subsets of topics that lead to improved system ranking estimation performance. In this section, we make a first step in that direction.

As *RS* proved overall to be the best performing algorithm in Section 5.5.1, we focus on it now. Recall, that *RS* is based on document popularity, that is, the most often retrieved documents have the highest chance of being sampled from the pool and thus being declared pseudo-relevant. This approach therefore assumes that *popularity* \approx *relevance*. It is clear, that this assumption is not realistic, but we can imagine cases of topics where it holds: in the case of *easy* topics. Easy topics are those where all or most systems do reasonably well, that is, they retrieve the truly relevant document towards the top of the ranking and then relevance can be approximated by popularity.

The above observation leads to the basic strategy we employ: adding topics to the set of topics according to their estimated difficulty. Again, as we do not have access to relevance judgments, we have to rely on an estimate of *collection topic hardness* [7] (see Figure 1.1, EA1), as provided by the Jensen-Shannon Divergence (*JSD*) approach by Aslam and Pavlu [7]. The *JSD* approach estimates a topic’s difficulty with respect to the collection and in the process also relies on different retrieval systems: the more diverse the result lists of different retrieval systems as measured by the Jensen-Shannon Divergence, the more difficult the topic is with respect to the collection.

Therefore, we perform a prediction task on two levels. First, a ranking of topics according to their inherent difficulty is estimated by the *JSD* approach and then we rely on the topics that have been predicted to be the easiest, to perform system ranking estimation. The only parameter of the *JSD* approach is the document cutoff. We relied on the parameter settings recommended in [7], that is a cutoff of 100 documents for TB-{04-06} and a cutoff of 20 documents for the remaining data sets.

The results of this two-level prediction approach are presented in Table 5.5.

Shown are Kendall’s τ on the full set of topics and the correlation achieved by *JSD* based selected subsets of $c = 10$ and $c = 20$ of topics. The particular size of the topic subset is of no great importance as seen in the small variation in τ . For nine out of sixteen data sets, we can observe improvements in correlation, though none of the improvements are statistically significant. The largest improvement in correlation is observed for EEXP-05, where the correlation on the full topic set $\tau = 0.48$ increases to $\tau = 0.62$ when the easiest $c = 10$ topics are evaluated. The correlation change of the data sets that degrade with *JSD* based topic subset selection is usually slight, the most poorly performing data set is NP-02, where $\tau = 0.69$ on the full set of topics degrades to $\tau = 0.60$. Considering the potential amount of improvements with the “right” subsets of topics as evident in Section 5.5.3 and Table 5.4, this result is somewhat disappointing. We suspect two reasons for the low levels of change the *JSD* approach achieves: apart from the fact, that the median AP strategy does in a few instances only perform little better than the baseline correlation (Section 5.5.3), the *JSD* approach itself does not estimate the topic’s difficulty to a very high degree. When evaluating the accuracy of the collection topic hardness results (Figure 1.1, EA1), *JSD* reaches correlations between $\tau = 0.41$ and 0.63 , depending on the data set.

We also evaluated three further automatic topic subset selection mechanisms. First, we attempted to exploit general knowledge we have about the performance of a number of retrieval approaches such as the fact that TFIDF usually performs worse than BM25 or Language Modeling. In order to rank n systems, we assumed to have an additional $k \ll n$ systems available for which we know the performance ranking based on past experience. The ranking of the $n+k$ systems is then estimated as usual, and, topic subset selection is performed by first selecting those topics, for which the k systems are ranked according to our assumption. The hypothesis being, that topics for which the system ranking estimator is able to derive the ranking of the known k system correctly, are more likely to also produce good estimates for the unknown n systems.

A second strategy is to cluster the estimated rankings derived for each topic and then to choose all topics of the largest cluster as topic subset. The motivation behind this approach can be explained by the results of Figure 5.3. We took the 10000 random subsets created for topic subsets of cardinality $c = 10$ in the TREC-7 data set and the *RS* approach. We sorted the subset samples according to the correlation they achieve with respect to the ground truth ranking, that is the MAP on the full set of $m = 50$ topics. Then, we created two sets: the set of *good* subsets, that are the 250 subsets with the highest correlation and the set of *bad* subsets, that are the 250 subsets with the lowest correlation with the ground truth. Now, for each of the 50 topics in the full topic set, we determined how often it appears in the *good* and *bad* sets. The ratio $|G|/(|G| + |B|)$ is 1 if a topic only appears in good sets, while it is 0 if a topic only appears in bad sets, a value of 0.5 means that the topic appears to the same extent in both types. In the plot in Figure 5.3 each point represents one topic. Two variations are given: the topics and their correlation to the MAP based ground truth and the topics and their correlation to the AP based ground truth. It is evident, that the best subsets are made out of topics which achieve a

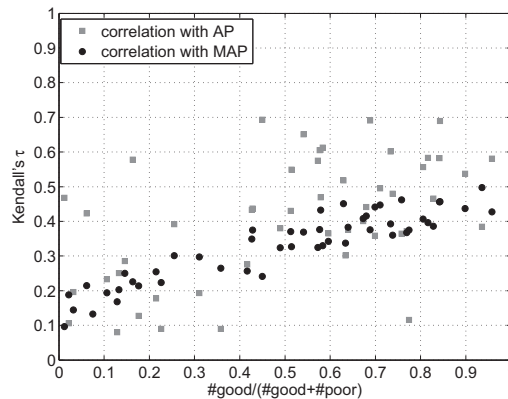


Figure 5.3: Distribution of topics in the best and worst subsets (RS , TREC-7): Of the 10000 random samples of topic subsets with cardinality $c = 10$, the best and worst performing 250 subsets are kept. For each of the 50 topics in TREC-7, it is recorded how often it appears in the good and bad subsets, the x-axis contains the ratio. Shown are the topics and their correlation with AP and their correlation with MAP.

high correlation with MAP. The topics that are predicted to the highest degree (their correlation with AP) are not necessarily those that are always found in the best sets – there are inter-relations with the other topics in the subset. As the best subsets are made out of topics that estimate rankings with a good correlation to the MAP based ground truth, we clustered the estimated rankings and chose as subset the cluster with the largest number of topics, estimating that this might be a cluster of good topics.

Finally, we attempted to approximate the estimation accuracy of a topic (Section 5.5.2) by introducing noise to the estimated performance scores and testing the robustness of the ranking against randomly introduced perturbations. If one estimator estimates the performance scores of three documents to be $(0.2, 0.201, 0.22)$ while a second estimator derives the scores $(0.2, 0.6, 1.0)$ we would have more confidence in the ranking of documents by the second estimator, as the score differences are larger, whereas the confidence in the document ranking is smaller for the first estimator, as the estimated performance scores are very similar. We tested this intuition by adding Gaussian noise to the estimated performances scores and determining by how much the ranking of documents after the introduction of noise differs from the ranking of documents based on the unperturbed scores. This method was motivated by the query performance prediction methods that work on query and document perturbations (Chapter 3).

The evaluation of those three more advanced strategies, however, failed to achieve better results than the JSD strategy. The reasons for the failure to identify valuable topic subsets with either of these mechanisms are not well understood yet and require further investigation.

| | RS | JSD | |
|-----------------|----------|--------------|--------------|
| | full set | c=10 | c=20 |
| TREC-6 | 0.443 | 0.455 | 0.485 |
| TREC-7 | 0.466 | 0.489 | 0.505 |
| TREC-8 | 0.538 | 0.585 | 0.588 |
| TREC-9 | 0.677 | 0.649 | 0.644 |
| TREC-10 | 0.643 | 0.634 | 0.635 |
| TB-04 | 0.708 | 0.760 | 0.733 |
| TB-05 | 0.659 | 0.670 | 0.612 |
| TB-06 | 0.518 | 0.495 | 0.508 |
| CLIR-01 | 0.702 | 0.706 | 0.698 |
| NP-02 | 0.693 | 0.623 | 0.597 |
| EDISC-05 | 0.666 | 0.709 | 0.729 |
| EEXP-05 | 0.483 | 0.616 | 0.616 |
| BLTR-06 | 0.523 | 0.501 | 0.528 |
| GEN-07 | 0.563 | 0.530 | 0.556 |
| LEGAL-07 | 0.741 | 0.695 | 0.728 |
| RELFB-08 | 0.559 | 0.589 | 0.638 |

Table 5.5: Overview of Kendall’s τ achieved by *RS* on the full set of topics and on topic subsets of cardinality $c = 10$ and $c = 20$ of the *JSD* topic subset selection strategy. In bold, improvements over the full topic set. All correlations reported are significant ($p < 0.005$), though none are statistically significantly different from the highest correlation per data set.

5.6 Conclusions

In this chapter, we have investigated the task of system ranking estimation, which attempts to rank a set of retrieval systems, for a given topic set and test corpus, according to their relative performance *without* relying on relevance judgments. This type of automatic evaluation could in the ideal case be used in the context of formal evaluations though currently the results suggest that this is not a realistic goal yet. We have described the most common approaches and performed an evaluation of them on a wider variety of data sets than done previously. In contrast to earlier findings [9, 114, 133, 135, 161] on a small number of older TREC data sets, we found the initially proposed approach by Soboroff et al. [133] to be the most stable and the best performing one. Moreover, we found the commonly reported problem of system ranking estimation methods, namely the severe underestimation of the performance of the best systems, not to be an inherent problem of system ranking estimation approaches. Instead we argue that this is a data set dependent issue, in particular it depends on the amount of human intervention in the best systems of a data set. If the best system is automatic in nature, or is derived with a small amount of human intervention, it can often be identified with a high degree of accuracy, or for some data sets, even correctly. This result suggests, that especially in practical applications, where we have a choice of different retrieval approaches it can be possible to automatically determine the best (or close to the best) performing one.

In terms of evaluation, it also proved beneficial to report not only the rank correlation coefficient Kendall's τ as evaluation measure of system ranking estimation approaches, but also to report the estimated rank of the best system as this measure provides an alternative view of an approach's performance.

In a second set of experiments, we turned to investigating the ability of system ranking estimation approaches to estimate the ranking of systems for each individual topic. We showed that the quality of the estimated rankings vary widely within a topic set. Based on this result, we designed a number of motivational experiments with different subset selection strategies. We were able to confirm the hypothesis that there exist subsets of topics that are better suited for the system ranking estimation task than others. Having found this regularity is only the first step however, for this knowledge to be useful in a practical setting, automatic methods are required that can identify those good subsets of topics to rely on.

We also proposed a strategy to automatically identify good subsets of topics by relying on topics that have been estimated to be easy. This strategy yielded some improvements, though they were not consistent across all data sets. Considering the amount of potential improvement, this can only be considered as a first attempt at subset selection.