A Case for Automatic System Evaluation

Claudia Hauff[1], Djoerd Hiemstra[1], Leif Azzopardi[2], and Franciska de Jong[1]

[1] University of Twente, Enschede, The Netherlands
{c.hauff,hiemstra,f.m.g.dejong}@ewi.utwente.nl
[2] University of Glasgow, Glasgow, UK
leif@dcs.gla.ac.uk

Abstract. Ranking a set retrieval systems according to their retrieval effectiveness without relying on relevance judgments was first explored by Soboroff *et al.* [13]. Over the years, a number of alternative approaches have been proposed, all of which have been evaluated on early TREC test collections. In this work, we perform a wider analysis of *system ranking estimation* methods on sixteen TREC data sets which cover more tasks and corpora than previously. Our analysis reveals that the performance of system ranking estimation approaches varies across topics. This observation motivates the hypothesis that the performance of such methods can be improved by selecting the "right" subset of topics from a topic set. We show that using topic subsets improves the performance of automatic system ranking methods by 26% on average, with a maximum of 60%. We also observe that the commonly experienced problem of underestimating the performance of the best systems is data set dependent and not inherent to system ranking estimation. These findings support the case for automatic system evaluation and motivate further research.

1 Introduction

Ranking retrieval systems according to their retrieval effectiveness *without* relying on costly relevance judgments was first explored by Soboroff *et al.* [13]. The motivation for this research stems from the high costs involved in the creation of test collections and in particular human relevance assessments. If system evaluation without relevance assessments is achievable, then the cost of evaluation could be greatly reduced. Automatic system evaluation could also be useful in the development of methods for data fusion and source selection.

In recent years, a number of *system ranking estimation* approaches have been proposed [5,12,13,14,16], which attempt to rank a set of retrieval systems (for a given topic set and a test corpus) without human relevance judgments. All approaches estimate a performance-based ranking of systems by considering the relationship of the top retrieved documents across systems. While the initial results highlighted the promise of this new direction, the utility of these approaches remains unclear. This is mainly because they usually underestimate the performance of the best systems, which is attributed to the "tyranny of the masses" effect [5]. In the analysis presented in this paper, we will show this

C. Gurrin et al. (Eds.): ECIR 2010, LNCS 5993, pp. 153–165, 2010.

problem not to be inherent to system ranking estimation methods. In previous work [5,12,13,14,16], the evaluations were mostly performed on the TREC-$\{3,5,6,7,8\}$ data sets[1]. In this work, perform a much wider analysis. We consider sixteen different TREC data sets, including a range of non-adhoc task data sets (such as expert search) and adhoc tasks on non-traditional corpora (such as the blog corpus). We find 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 manual runs of a data set[2].

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 [11] it was found that some topics are better suited than others to differentiate the performance of retrieval systems. We hypothesize and verify experimentally, that with the right subset of topics, the current methods for estimating system rankings without relevance judgment can be significantly improved. These findings suggest that under certain conditions, automatic system evaluation is a viable alternative to human relevance-judgments based evaluations.

This paper is organized as follows: in Sec. 2 we provide an overview of related work and the motivation for relying on subsets of topics for system ranking. The research questions and the experimental setup are outlined in Sec. 3. The result section (Sec. 4) contains (i) a comparison of four ranking estimation approaches, (ii) motivational experiments to show the validity of using topic subsets, and, (iii) a first attempt to automatically find "good" subsets of topics. In Sec. 5 conclusions are drawn and directions for future work are discussed.

2 Related Work

Research aiming to reduce the cost of evaluation has been conducted along two lines: a number of approaches focus on *reducing* the amount of manual assessments required [2,6,8], while others rely on *fully automatic* evaluation. We only consider approaches of the second category, that is, we focus on algorithms that require no manual assessments at all.

The first work in this area is attributed to Soboroff *et al.* [13]. It was motivated by the fact that the relative ranking of retrieval systems remains largely unaffected by the assessor disagreement in the creation of relevance judgments [15]. This observation led to the proposal to use automatically created *pseudo* relevance judgments which are derived as follows: first, the top retrieved documents of all systems to rank for a topic are pooled together such that a document that

[1] When we speak of a data set, such as TREC-3, we mean all retrieval runs submitted to the Text REtrieval Conference (TREC, http://trec.nist.gov/) for the topics of that task. A retrieval run is the output of a retrieval system and thus by ranking retrieval runs, we rank retrieval systems.

[2] 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 (e.g. by providing explicit relevance judgments, manually re-ranking documents etc.).

is retrieved by x systems, appears x times in the pool. Then, a number of documents, the so called *pseudo relevant documents*, are drawn at random from the pool. This process is performed for each topic and the subsequent evaluation of each system is performed with pseudo relevance judgments in place of relevance judgments. A system's effectiveness is estimated by its pseudo mean average precision. To determine the accuracy of the estimated system ranking, it is compared against the ground truth ranking, that is the ranking of systems according to Mean Average Precision (MAP). The experiments in [13] were performed on TREC-{3,5,6,7,8}. The reported correlations were significant, however, one major drawback was discovered: whereas the ranking of the poorly and moderately performing systems was estimated quite accurately, the best performing systems were always ranked too low. It was suggested in [5] that this observation can be explained by the "tyranny of the masses" effect, where the best systems are estimated to perform poorly due to being different from the average system.

The exploitation of pseudo relevant documents has been further investigated by Nuray & Can [12], on very similar data sets (TREC-{3,5,6,7}). In contrast to [13], not all systems to be ranked participate in the derivation of pseudo relevance judgments. The authors experimented with different methods to find a good subset of $P\%$ of systems; overall, the best approach was to select those systems that were most different from the average system. Once a subset of non-average 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. The best performing result list merging mechanism was found to be Condorcet voting, where each document in the list is assigned a value according to its rank. This way, not only the frequency of occurrence of a document in various result lists is a factor as in [13], but also the rank the document is retrieved at. The reported correlations were generally higher than in [13]. However, our experiments will show that this is not always the case when evaluating a wider range of data sets.

In [16] it was proposed to rank the 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. Experiments on TREC-{3,5,6,7,10} generally yielded lower correlations than in [5,12,13]. A somewhat similar strategy, the structure of overlap method, was proposed by Spoerri [14]. In contrast to [5,12,13,16], not all systems are ranked at once, instead random groupings of five systems are ranked repeatedly. For each grouping and for each of the topics, the percentage $S\%$ of documents in the ranked lists found by only one and the percentage $A\%$ of documents found by all five systems are determined. The three scores $S\%$, $A\%$ and $(S\% - A\%)$ were proposed as estimated system score. These scores are further averaged across the number of topics in the topic set. Since each system participates in a number of groupings, the scores across those groupings are again averaged, leading to the final system score. The reported correlations are significantly higher than in [5,12,13,16]. However, not all the systems that participated in TREC were used, only particular subsets (specifically the best automatic systems). In Sec. 4, we will show that this approach does not perform better than the originally proposed method by Soboroff *et al.* [13] when all available systems are included.

While each system ranking estimation method takes a different approach, a common underlying assumption is that all topics are equally useful when estimating the performance based 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 [11]. In order to explore the relationship between a set of topics and a set of systems, Mizzaro & Robertson [11] 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. While the study in [11] was theoretical in nature, an empirical study has been performed by Guiver *et al.* [9] yielding similar conclusions: that selecting the right subset of topics provides a similar indication of relative system performance to using the full topic set. If the right subset of topics could be selected, then the number of topics needed to compare systems could be greatly reduced. In this work, we examine whether topic subsets could also be used to improve the performance of system ranking estimation methods.

3 An Analysis of System Ranking Estimation Methods

Given the prior research on automatic system ranking estimation, the following analysis of these methods is undertaken to (i) evaluate the main approaches on a wide variety of data sets, (ii) to validate (or not) previous findings, and (iii) to improve the current approaches by using subsets of the full set of topics. Specifically, we examine the following research questions:

1. 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?
2. By reducing the topic set size, can the performance of current system ranking estimation methods be improved?
3. Can topic subsets, that improve the performance of system ranking estimation methods, be selected automatically?

We focus the analysis on four system ranking estimation approaches [7,12,13,14] and evaluate them across sixteen different data sets. The following subsections detail the data sets used, the system ranking estimation approaches and the commonly employed evaluation measure of system ranking estimation.

3.1 Data Sets

As in [5,12,13,14,16], we rely on TREC tasks over different years in our experiments. However, whereas earlier studies focused mainly on TREC-{3,5,6,7,8}, we include a much wider variety of data sets. In particular, we evaluated **TREC-{6,7,8}** (adhoc tasks on the TREC Vol. 4+5 corpus), **TREC-{9,10}** (adhoc tasks on the WT10g corpus), **TB-{04,05,06}** (adhoc tasks on the GOV2 corpus with topics from the TeraByte tracks), **CLIR-01** (Cross-Language track 2001), **NP-02** (Named Page Finding track 2001), **EDISC-05** (Enterprise Discussion track 2005), **EEXP-05** (Enterprise Expert Search track 2005), **BLTR-06** (Blog Topical Relevance track 2006), **GEN-07** (Genomics track 2007), **LEGAL-07** (Legal track 2007) and **RELFB-08** (Relevance Feedback track 2008). All data sets can be downloaded from the TREC website.

The number of retrieval systems to rank for each data set varies between 37 and 129, while the number of topics, that are used to evaluate the systems, ranges from 25 to 237 (Tab. 1). Included in our experiments are all available runs, automatic as well as manual and short as well as long runs. We preprocessed the available corpora (TREC Vol. 4+5, WT10g and GOV2) by applying Krovetz stemming [10] and stopword removal.

3.2 Algorithms

Based on the results in the literature, we employ four different system ranking estimation methods: the data fusion (*DF*) approach by Nuray & Can [12], the random sampling (*RS*) approach by Soboroff *et al.* [13], the structure of overlap approach (*SO*) by Spoerri [14], and the document similarity auto-correlation (*ACSim*) approach by Diaz [7].

While *DF*, *RS* and *SO* were introduced in Sec. 2, *ACSim* has not been applied to system ranking estimation yet. The motivation for evaluating these approaches is their mix of information sources. *RS* relies on document overlap as shown in [5], while *DF* takes the rank a system assigns to a document into account. *ACSim* goes a step further and considers the content similarity of documents. Finally, *SO* ranks a large number of subsets of systems to achieve a ranking across all systems. Due to space constraints, we only briefly sketch the parameter settings of *DF*, *RS* and *SO* as these methods have already been described in Sec. 2.

Data fusion (*DF*). We evaluate the three parameters of the approach over a wide range of values: $s = \{1\%, 5\%, 10\%, 20\%, .., 50\%\}$, $P = \{10\%, 20\%, .., 100\%\}$ and $b = \{10, 20, .., 100, 125, .., 250\}$. To determine the parameter setting of a data set, we train on the remaining data sets available for that corpus, e.g. the parameters of TREC-6 are those that lead to the best performance on TREC-{7,8}. Data sets for training are only available for TREC-{6-10} and TB-{04-06} though. For the remaining data sets, we choose the parameter setting, that gives the best performance across those data sets: $s = 10\%$, $b = 50$ and $P = 100\%$, that is, the best results are achieved when *not* biasing the selection of systems ($P = 100\%$). Since the parameters are optimized on the test set, we expect *DF* to perform very well in those instances.

Random sampling (*RS*). We follow the methodology from [13] and rely on the 100 top retrieved documents per system. 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.

This requires some knowledge about the distribution of relevance judgments; this proves not to be problematic however, as fixing the percentage to a small value yields little variation in the results. As in [13], due to the randomness of the process, we perform 50 trials and average the results.

Structure of overlap (*SO*). As in [14], we rely on the top 50 retrieved documents per system and report the results of the $(S\% - A\%)$ score, as it gives the best results which is in accordance to [14].

Document similarity auto-correlation (*ACSim*). *ACSim* [7][3], 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 retrieval scores vector $\mathbf{y}$ of the top 75 retrieved documents, a perturbed score vector $\tilde{\mathbf{y}}$ is derived. Each element y_i is replaced in $\tilde{\mathbf{y}}$ by the weighted average of scores of the 5 most similar documents (based on TF.IDF) in $\mathbf{y}$. If the cluster hypothesis is fullfilled, we expect the most similar documents to also receive a similar score by the retrieval system, otherwise high document similarity is not expressed in similar scores and $\tilde{y}_i$ will be different from y_i. The average score vector $\mathbf{y}_\mu$ is formed by averaging $\mathbf{y}$ over all systems. To score each system, the linear correlation coefficient between $\mathbf{y}_\mu$ and $\tilde{\mathbf{y}}$ is determined. Here, a system is estimated to perform well, if it is similar to the average system. In [7] this approach has been used to rank systems according to single topics, while we use it to rank systems across a set of topics. Please note, that we can only report *ACSim* for the data sets, for which the corpora were available to us: TREC-{6-10} and TB-{04-06}.

3.3 Correlation of System Rankings

Each system ranking estimation method outputs a list of systems ordered by estimated system performance (i.e. system ranking). To provide an indication of the quality of an approach, this ranking is correlated against the ground truth ranking, which is determined by the retrieval effectiveness using relevance judgments. In all but two data sets, the effectiveness measure is MAP. Mean reciprocal rank and statistical AP [4] are the measures used for NP-02 and RELFB-08 respectively. Reported is the rank correlation coefficient Kendall's Tau $\tau \in [-1, 1]$, which measures the degree of correspondence between two rankings [1].

4 Empirical Findings

In Sec. 4.1 we compare the performances of *DF*, *RS*, *SO* and *ACSim* on the full set of topics. Then, in Sec. 4.2, we will show that: (i) system rankings cannot be estimated equally well for each topic in a topic set, (ii) relying on topic subsets can improve the accuracy of system ranking estimation, and (iii) automatically selecting subsets of topics can lead to improvements in system ranking.

[3] Referred to as $\rho(\tilde{\mathbf{y}}, \mathbf{y}_\mu)$ in [7].

4.1 System Ranking Estimation on the Full Set of Topics

In Tab. 1, the results of the evaluation are shown in terms of Kendall's τ. *DF* performs best on TREC-{6,7}, the poor result on TREC-8 is due to an extreme parameter setting learned from TREC-{6,7}. The highly data set dependent behavior of *DF* is due to its bias: the pseudo relevant documents are selected from non-average systems. A system that is dissimilar to the average system, can either perform very well or very poorly. *RS* outperforms *DF* on the remaining six data sets, where training data for *DF* is available (TREC-{9,10}, TB-{0,4-06}). On the eight data sets without training data, *DF*'s parameters were optimized on the test set - despite this optimization *RS* outperforms *DF* in four instances.

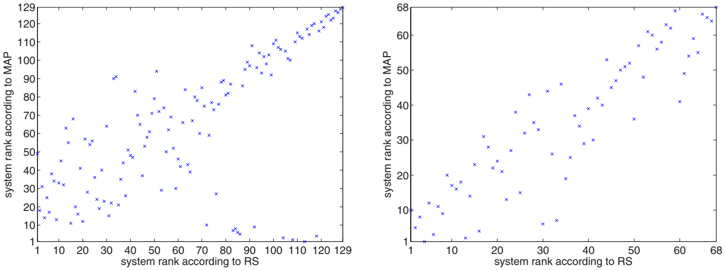
Table 1. System ranking estimation on the full set of topics. All correlations reported are significant ($p < 0.005$). The highest τ per data set is bold. Column **#sys** contains the number of retrieval systems to rank for a data set with **#top** topics. The final three columns contain the name of the best system according to the ground truth which is expressed in mean reciprocal rank (NP-02), statistical AP [4] (RELFB-08) and MAP (all other data sets) respectively. **M/A** indicates whether the best system is manual (M) or automatic (A). **ER** shows the estimated rank of the best system by the *RS* approach. Rank 1 is the top rank.

			Kendall's Tau τ				**RS based rank estimate**		
	#sys	**#top**	**DF**	**ACSim**	**SO**	**RS**	**best system**	**M/A**	**ER**
TREC-6	73	50	**0.600**	0.425	0.470	0.443	*uwmt6a0*	M	57
TREC-7	103	50	**0.486**	0.417	0.463	0.466	*CLARIT98COMB*	M	74
TREC-8	129	50	0.395	0.467	0.532	**0.538**	*READWARE2*	M	113
TREC-9	105	50	0.527	0.639	0.634	**0.677**	*iit00m*	M	76
TREC-10	97	50	0.621	**0.649**	0.598	0.643	*iit01m*	M	83
TB-04	70	50	0.584	0.647	0.614	**0.708**	*uogTBQEL*	A	30
TB-05	58	50	0.606	0.574	0.604	**0.659**	*indri05AdmfL*	A	32
TB-06	80	50	0.513	0.458	0.447	**0.518**	*indri06AtdnD*	A	20
CLIR-01	47	25	0.697	-	0.650	**0.702**	*BBN10XLB*	A	2
NP-02	70	150	0.667	-	0.668	**0.693**	*thunp3*	A	17
EDISC-05	57	59	**0.668**	-	0.614	0.666	*TITLETRANS*	A	1
EEXP-05	37	50	**0.589**	-	0.502	0.483	*THUENT0505*	A	10
BLTR-06	56	50	0.482	-	0.357	**0.523**	*wxoqf2*	A	5
GEN-07	66	36	**0.578**	-	0.362	0.563	*NLMinter*	M	1
LEGAL-07	68	43	**0.754**	-	0.749	0.741	*otL07frw*	M	4
RELFB-08	117	237	0.537	-	0.544	**0.559**	*Brown.E1*	M	65

Relying on TF.IDF based document similarity does not aid, shown by *ACSim* performing worse than *RS* in seven out of eight data sets. *SO* performs slightly better than *RS* on three data sets, on the remaining twelve its performance is worse.

As discussed, the commonly cited problem of automatic system evaluation is the mis-ranking of the best systems. To give a better impression of the ranking accuracy, in Fig. 1 scatter plots of the estimated system ranks versus the MAP based ground truth system ranks are shown for two data sets. Apart from the best systems, which are severely mis-ranked in Fig. 1(a), the estimated ranking shows a good correspondence to the true ranking. In contrast, in Fig. 1(b), the best systems are estimated more accurately.

As in previous work evaluations have mostly been carried out on early TREC data sets, where the problem of underestimating the best systems occurs



(a) TREC-8, RS, $\tau = 0.54$ (b) LEGAL-07, RS, $\tau = 0.74$

Fig. 1. Scatter plots of estimated system ranks versus ground truth (MAP) based system ranks. Each marker indicates one system. Rank 1 is assigned to the best system.

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. Tab. 1 also contains the estimated rank (ER) of the best system by the *RS* approach, which in the ground truth is assigned rank 1. For instance, while for TREC-8, the best system is estimated to be ranked at rank 113 ($\tau = 0.54$), the corresponding estimated ranking on GEN-07 is 1 ($\tau = 0.56$), that is, the best system is correctly identified. The fact that both τ values are similar, suggests that both, τ and ER, should be reported, as sometimes we are most interested in identifying the best systems correctly. A look at the best systems for each data set shows, that in TREC-{6-10} the best systems are manual and derived with a great deal of human intervention[4]. The best systems of TB-{04-06} on the other hand, are automatic. A similar observation can be made for the more recent data sets, the best systems are mostly automatic. In the case of GEN-07 and LEGAL-07, where the best systems are classified as manual, the human intervention is less pronounced[5]. The severe mis-ranking of RELFB-08's best system is a result of the task, which is to exploit manually judged documents for ranking, that is the systems are based on considerable human intervention.

We conclude, that in contrast to previous work, *RS* is the most consistent and overall the best performing approach. Furthermore, in contrast to common belief, automatic system evaluation methods are capable of identifying the best runs, when they are automatic or involve little human intervention. The extent of the mis-ranking problem is largely influenced by the amount of human intervention in the best manual runs.

4.2 System Ranking Estimation on Topic Subsets

In this section, we will show that the ability of system ranking estimation methods to rank systems correctly, differs significantly between the topics of a topic

[4] E.g. the best system of TREC-6, *uwmt6a0*, was created by letting four human assessors judge documents for their relevance for more than 100 hours.

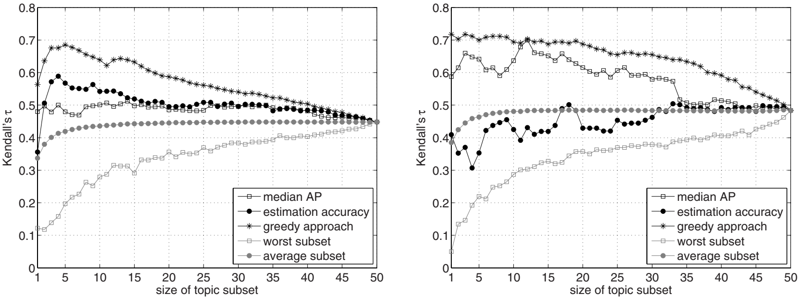
[5] E.g. the best system of GEN-07, *NLMinter*, relied on manually reformulated queries.

set. While for a number of topics the estimated rankings are highly accurate and close to the ground truth rankings, for other topics they fail. Based on this observation, we hypothesize that the performance of system ranking estimation approaches can be improved, if the "right" subset of topics is used.

Single Topic Performance. For each topic, we evaluate the estimated ranking of systems by correlating it against the ground truth ranking which is based on average precision. 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. We are interested in how well the system ranking estimation approach performs for each individual topic. Due to space constraints, in Tab. 2, we only present results for four data sets. Across the data sets and system ranking estimation methods, the spread in correlation between the best and worst case is very wide; in the worst case, there is no significant correlation between the ground truth and the estimated ranking, in the best case the estimated ranking is highly accurate. These findings form our motivation: if we can determine a subset of topics for which the system ranking estimation algorithms perform well, we hypothesize that this will enable us to achieve a higher estimation accuracy of the true ranking across the full set of topics.

Table 2. Single topic dependent ranking performance: minimum and maximum estimation ranking accuracy in terms of Kendall's τ. Significant correlations ($p < 0.005$) are marked with †.

	DF		**ACSim**		**SO**		**RS**	
	min. τ	**max. τ**	**min. τ**	**max.**	**min. τ**	**max. τ**	**min. τ**	**max. τ**
TREC-6	0.008	0.849†	−0.134	0.777†	−0.147	0.752†	−0.106	0.823†
TB-04	0.002	0.906†	−0.038	0.704†	−0.056	0.784†	−0.025	0.882†
CLIR-01	0.268	0.862†	-	-	0.221	0.876†	0.248	0.839†
LEGAL-07	0.027	0.690†	-	-	0.058	0.691†	−0.008	0.690†



(a) *RS*, TREC-6 (b) *RS*, EEXP-05

Fig. 2. Topic subset selection experiments

Topic Subset Performance. Having shown that the quality of ranking estimation varies between topics, we now investigate if selecting a subset of topics from the full topic set is useful in the context of system ranking estimation

Table 3. Topic subset selection experiments: comparison of Kendall's τ achieved by *RS* when relying on the full topic set and the best greedy subset, ± indicates the percentage of change. All τ values are significant ($p < 0.005$). The *JSD* columns show the results when performing automatic topic subset selection with the *JSD* approach. Reported are the results of topic subset sizes $c = 10$ and $c = 20$. In bold, improvements of *JSD* over the full topic set.

	RS			**JSD**	
	full set	**greedy**	±%	**c=10**	**c=20**
	τ	τ		τ	τ
TREC-6	0.443	0.654	+47.6%	**0.455**	**0.485**
TREC-7	0.466	0.584	+25.3%	**0.489**	**0.505**
TREC-8	0.538	0.648	+20.4%	**0.585**	**0.588**
TREC-9	0.677	0.779	+15.1%	0.649	0.644
TREC-10	0.643	0.734	+14.2%	0.634	0.635
TB-04	0.708	0.846	+19.5%	**0.760**	**0.733**
TB-05	0.659	0.812	+23.2%	**0.670**	0.612
TB-06	0.518	0.704	+35.9%	0.495	0.508
CLIR-01	0.702	0.808	+15.1%	**0.706**	0.698
NP-02	0.693	0.853	+23.1%	0.623	0.597
EDISC-05	0.666	0.801	+20.3%	**0.709**	**0.729**
EEXP-05	0.483	0.718	+48.7%	**0.616**	**0.616**
BLTR-06	0.523	0.601	+14.9%	0.501	**0.528**
GEN-07	0.563	0.678	+20.4%	0.530	0.556
LEGAL-07	0.741	0.865	+16.7%	0.695	0.728
RELFB-08	0.559	0.872	+60.0%	**0.589**	**0.638**

algorithms. That is, we try to determine whether we can improve the accuracy of the algorithms over the results reported in Sec. 4.1 on the full set of topics. To investigate this point, we experiment with selecting subsets of topics according to different strategies.

A topic set consists of m topics. We thus test topic subsets of cardinality $c = \{1, ..., m\}$. Ideally, for each c we would test all possible subsets. As this is not feasible[6], for each c, we randomly sample 10000 subsets of topics, run a system ranking estimation algorithm, evaluate it and record the (i) the worst τ and (ii) the average τ achieved across all samples[7]. We also include three iterative subset selection strategies: (iii) the greedy strategy, where a subset of topics is greedily built by adding one topic at a time such that τ is maximized, (iv) the median AP strategy where a subset of topics is built by each time adding the easiest topic, which is the topic that exhibits the highest median average precision across all systems and (v) the estimation accuracy strategy where a subset of topics is built by each time adding the topic with the highest estimation accuracy (single topic performance). As strategies (iii)-(v) require knowledge of the relevance judgments, these experiments should only be seen as motivational: the goal is to determine whether it is beneficial at all to rely on subsets instead of the full topic set.

For the topic subsets of each cardinality, we determine the correlation between the estimated ranking of systems and the ground truth ranking of systems across the full set of topics. Now we are indeed interested in how well a subset of one or

[6] For each cardinality c, a total of $\binom{m}{c}$ different subsets exist.

[7] Recording the best τ of the sampled subsets does not perform better than the greedy approach and is therefore not reported separately.

more topics can be used to approximate the ranking of systems over the entire topic set.

Indicatively for two data sets, the results are shown in Fig. 2. The results of the other data sets and algorithms not shown are similar. The greedy strategy, especially at low subset sizes, yields significantly higher correlations than the baseline, which is the correlation at the full topic set size. The worst subset strategy shows the potential danger of choosing the wrong subset of topics - τ is significantly lower than the baseline for small c. 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. For the median AP strategy, where first the easiest topics are added to the subset of topics, the gains in correlation over the baseline are visible, though less pronounced than for the greedy strategy. Less consistent is the estimation accuracy strategy, where first those topics are added to the topic subset, for whom the ranking of systems is estimated most accurately. While in Fig. 2(a) this strategy comes closest to the greedy approach, in Fig. 2(b) this strategy most of the time performs worse than the average τ strategy.

In Tab. 3, a summary of the results of the *RS* approach across all data sets is given (the results of *DF*, *SO*, *ACSim* are similar but not shown due to lack of space): shown is Kendall's τ achieved on the full set of topics and the best performing τ of the greedy approach. The percentage of change varies between 14.2% and 60% with a maximum τ of 0.872. Thus, across all data sets, subsets of topics indeed exist that can significantly improve the accuracy of system ranking estimation.

Automatic Topic Subset Selection. Topic subset selection will only be useful in practice, if it becomes possible to automatically identify useful subsets. As *RS* proved overall to perform best, we focus on it. *RS* is popularity based, 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 assumes that *popularity* $\approx$ *relevance*. 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 results of the median AP strategy in Fig. 2 confirms this reasoning.

This leads to the following strategy: adding topics to the subset of topics according to their estimated difficulty. As we do not have access to relevance judgments, we rely on an estimate of topic difficulty, as provided by the Jensen-Shannon Divergence (JSD) approach [3]. 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 systems as measured by the JSD, the more difficult the topic is. Thus, we perform an estimation task on two levels: first, a ranking of topics according to their difficulty is estimated, then we rely on the topics estimated to be easiest for system ranking estimation.

The results of this automatic subset selection approach are shown in the last two columns of Tab. 3: listed are the τ values achieved for subsets of 10 and 20 topics estimated to be easiest. The particular size of the topic subset is of no great importance as seen in the small variation in τ. It is also visible that this approach can help, though the improvements are small and inconsistent. As reason we suspect the fact, that the accuracy of the JSD approach itself is also limited [3].

5 Summary and Future Work

In this work, we have compared four system ranking estimation approaches on a wide variety of data sets and, contrary to earlier findings, we have shown that the first approach proposed (Soboroff *et al.* [13]) is still the best performing method. We also observed that the ability of system ranking estimation methods varies widely for each individual topic. A number of experiments with different topic subset selection strategies confirmed the hypothesis that some subsets of topics are better suited for the system ranking estimation algorithms than others.

We 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 improvement, though not consistently across all data sets. Considering the potential for improvement, this result should be considered as a first attempt at topic subset selection. In the future, we will focus on identifying topic features that distinguish between topics appearing in subsets which improve system ranking estimation and topics that occur mostly in poorly performing subsets.

Finally, we made the important finding, that the problem of mis-ranking the best systems encountered by previous authors, which has so far hampered the take-up of automatic system ranking estimation, is strongly data set dependent. In particular we found that the smaller the amount is of human intervention in the best systems, the smaller the problem of underestimating their performance. In fact, for some TREC data sets we were indeed able to identify the best system correctly, thus making the case for automatic system evaluation.

References

1. Rank Correlation Methods. Hafner Publishing Co., New York (1955)
2. Amitay, E., Carmel, D., Lempel, R., Soffer, A.: Scaling ir-system evaluation using term relevance sets. In: SIGIR 2004, pp. 10–17 (2004)
3. Aslam, J.A., Pavlu, V.: Query hardness estimation using Jensen-Shannon divergence among multiple scoring functions. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECiR 2007. LNCS, vol. 4425, pp. 198–209. Springer, Heidelberg (2007)
4. Aslam, J.A., Pavlu, V., Yilmaz, E.: A statistical method for system evaluation using incomplete judgments. In: SIGIR 2006, pp. 541–548 (2006)
5. Aslam, J.A., Savell, R.: On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In: SIGIR 2003, pp. 361–362 (2003)
6. Carterette, B., Allan, J.: Incremental test collections. In: CIKM 2005, pp. 680–687 (2005)

7. Diaz, F.: Performance prediction using spatial autocorrelation. In: SIGIR 2007, pp. 583–590 (2007)
8. Efron, M.: Using multiple query aspects to build test collections without human relevance judgments. In: ECIR 2009, pp. 276–287 (2009)
9. Guiver, J., Mizzaro, S., Robertson, S.: A few good topics: Experiments in topic set reduction for retrieval evaluation. To appear in TOIS
10. Krovetz, R.: Viewing morphology as an inference process. In: SIGIR 1993, pp. 191–202 (1993)
11. Mizzaro, S., Robertson, S.: Hits hits trec: exploring ir evaluation results with network analysis. In: SIGIR 2007, pp. 479–486 (2007)
12. Nuray, R., Can, F.: Automatic ranking of information retrieval systems using data fusion. Information Processing and Management 42(3), 595–614 (2006)
13. Soboroff, I., Nicholas, C., Cahan, P.: Ranking retrieval systems without relevance judgments. In: SIGIR 2001, pp. 66–73 (2001)
14. Spoerri, A.: Using the structure of overlap between search results to rank retrieval systems without relevance judgments. Information Processing and Management 43(4), 1059–1070 (2007)
15. Voorhees, E.M.: Variations in relevance judgments and the measurement of retrieval effectiveness. Information Processing and Management 36, 697–716 (2000)
16. Wu, S., Crestani, F.: Methods for ranking information retrieval systems without relevance judgments. In: Matsui, M., Zuccherato, R.J. (eds.) SAC 2003. LNCS, vol. 3006, pp. 811–816. Springer, Heidelberg (2004)