Relying on Topic Subsets for System Ranking Estimation

Claudia Hauff, Djoerd Hiemstra, Franciska de Jong
University of Twente
Enschede, the Netherlands
{ c.hauff, hiemstra, f.m.g.dejong}@ewi.utwente.nl

Leif Azzopardi
University of Glasgow
Glasgow, United Kingdom
leif@dcs.gla.ac.uk

ABSTRACT

Ranking a number of retrieval systems according to their retrieval effectiveness without relying on costly relevance judgments was first explored by Soboroff *et al* [6]. Over the years, a number of alternative approaches have been proposed. We perform a comprehensive analysis of system ranking estimation approaches on a wide variety of TREC test collections and topics sets. Our analysis reveals that the performance of such approaches is highly dependent upon the topic or topic subset, used for estimation. We hypothesize that the performance of system ranking estimation approaches can be improved by selecting the "right" subset of topics and show that using topic subsets improves the performance by 32% on average, with a maximum improvement of up to 70% in some cases.

Categories and Subject Descriptors: H.3.4 Information Storage and Retrieval: Information Search and Retrieval

General Terms: Experimentation.

Keywords: Evaluation, System Ranking Estimation.

1. INTRODUCTION

Estimating the performance of retrieval systems without recourse to relevance judgments was first explored by Soboroff *et al* [6]. The motivation stems from the high costs involved in the creation of test collections. Moreover, in a dynamic environment such as the Web regular evaluation of search engines with manual assessments is not feasible [6]. In recent years, a number of *system ranking estimation* approaches have been proposed [1, 5, 6, 7, 8] that attempt to rank a set of retrieval systems without the need for manual relevance judgments. In each of the these approaches, the retrieval results of the full TREC topic set are relied upon to form an estimate of system performance. However, in [4] it was found that some topics of a topic set are better suited than others to differentiate the performance of retrieval systems. While the work in [4] was performed in a more general evaluation context, here we explore the application of this observation in the context of system ranking estimation. In this paper, we hypothesize, that if the "right" subset of topics (i.e. those that best differentiate the performance of retrieval systems) is used, the current methods for estimating system ranking without relevance judgment can be substantially and significantly improved. To this aim, we implemented four different approaches to system ranking estimation and compare their performance in a comprehensive analysis. We empirically determine the extent of the topic dependent performance and perform a range of experiments to evaluate the degree to which topic subsets can improve the performance of system ranking estimation approaches. In contrast to previous work, the evaluation is conducted on a wider variety of test collections, including more recent ones. Our results show that the quality of system ranking estimation methods varies considerably depending on the set of topics, and subset of topics.

The paper is organized as follows: first, in Sec. 2, a brief overview of related work is given. Then, in Sec. 3, we describe the motivation for our experiments. The experimental setup is outlined in Sec. 4, followed by the empirical analysis in Sec. 5. We conclude with a summary in Sec. 6.

CIKM'09, November 2–6, 2009, Hong Kong, China.

2. RELATED WORK

Soboroff *et al* [6] proposed to rely on automatically derived pseudo relevance judgments to rank retrieval systems instead of costly manual relevance assessments. For each topic, the top retrieved documents across the retrieval systems to rank are pooled. A number of documents are sampled from the pool and used as pseudo relevant documents. The subsequent evaluation of each system is then performed with pseudo relevance judgments in place of relevance judgments. Although the reported correlations with the ground truth, that is the ranking of systems based on mean average precision (MAP), were significant, the performances of the best systems were consistently underestimated. It was suggested in [1] that this observation is due to the best systems being too different from the average.

The exploitation of pseudo relevant documents was also investigated by Nuray & Can [5]. In contrast to [6], not all available systems participate in the creation of pseudo relevance judgments, only those that are the most different from the average. Additionally, the rank a document is retrieved at is taken into account. The reported results are generally higher than those reported in [6] for the topic sets evaluated. However, in this work, we perform an extensive evaluation across more test collections, and show that this approach does not always deliver better performance.

Other methods that estimate a ranking of systems based on document overlap such as [8, 7], have not been proven to be as successful as [6] in our experiments when considering all available systems for ranking.

The aforementioned methods have all assumed that all topics are 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 [4].

3. TOPIC SUBSET SELECTION

To explore the relationship between a set of topics and a set of systems, Mizarro & Robertson [4] 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, Mizarro & Robertson [4] employed HITS, a method that returns a hub and authority value for each node. While the study in [4] was more theoretic in nature, a recent follow up on this work by Guiver *et al* [3] showed 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 to reduce the number of topics required.

The finding that individual topics vary in their ability to indicate system performance provided the basis for our work as it implies that there might exist a subset of topics that is as suited to estimate system performance as the full set of topics. While the motivation in [3, 4] is to reduce the cost of evaluation by reducing the topic set size, we are motivated by the fact that system ranking estimation does not perform equally well across all topics.

We examine the following research question: By reducing the topic set size, can the performance of current system ranking estimation methods be improved?

4. DATA SETS AND ALGORITHMS

We conduct our analysis on eight different topic sets: TREC-{6,7,8} (TREC Volumes 4+5 minus CR corpus), TREC-{9,10} (WT10g corpus) and TB-{04,05,06} (GOV2 corpus)[1]. Each topic set contains 50 topics, the number of retrieval systems to rank varies between 58 and 129 (Table 1).

The estimation methods we evaluate are: the data fusion (DF) approach by Nuray & Can [5], the random sampling (RS) approach by Soboroff *et al* [6], the document similarity ($ACSim$) and the document score approach ($ACScore$) by Diaz [2]. The latter two approaches were originally applied to rank systems for a single topic, not across a set of topics. The main motivation for evaluating specifically those approaches is their mix of information sources. In particular, RS relies on document overlap, while DF and $ACScore$ take the rank and retrieval score, respectively, a system assigns to a document into account. Finally, the $ACSim$ approach goes a step further and considers the content similarity between ranked documents.

DF: The variation of the data fusion approach, that performed best in [5] and which we utilize (Condorcet voting and biased system selection), has three parameters to set. We determined each topic set's parameters by training on the remaining topic sets available for that corpus.

RS: We follow the methodology of [6] and rely on the 100 top retrieved documents per retrieval system. We pool the results of all systems that are to be ranked. As in [6], due to the inherent randomness of the process, we perform 50 trials. In the end, we average the pseudo AP values for each pair of topic and system.

ACScore: This approach, proposed in [2][2], is based on document overlap and the score, a retrieval system assigns to each document. Essentially, a retrieval system is estimated to perform well, if its documents' scores are close to the average scores across all systems.

ACSim: The second approach from [2][3] we evaluate, is based on the notion that well performing systems are likely to fulfill the cluster hypothesis, while poorly performing systems are not.

The evaluation is performed by reporting the rank correlation coefficient Kendall's Tau $\tau \in [-1, 1]$, between the ground truth ranking, that is the ranking of retrieval systems according to MAP, and the system ranking produced by the ranking estimation methods.

5. EXPERIMENTS

In Sec. 5.1, we compare the four ranking estimation approaches on the full topic sets. In Sec. 5.2, we will show that the system ranking cannot be estimated equally well for each topic of a topic set. Finally, in Sec. 5.3, we perform a number of motivational experiments to determine if it is possible to exploit this observation.

5.1 System Ranking Estimation

In Table 1, the results of the evaluation of the four system ranking estimation methods are shown. DF performs well on TREC-{6,7}, the poor result on TREC-8 is due to an extreme parameter setting found during training. RS on the other hand outperforms DF on both the topic sets of WT10g and GOV2. Relying on content similarity does not aid, shown by $ACSim$ performing worse than $ACScore$. Conversely, the knowledge of the retrieval scores assigned to a document by a system, as exploited by $ACScore$, leads to a slightly worse performance than RS, which considers document overlap only. If we consider the mean of Kendall's τ across all topic sets, RS shows the most consistent performance, followed by $ACScore$.

Of note is the high correlations that the four approaches achieve on the topic sets of WT10g in comparison to TREC Volumes 4+5. System ranking estimation is harder on TREC Volumes 4+5 due to the greater number of manual runs available for those topic sets. Manual runs are often very different from automatic runs. We also observe that the DF approach, which is designed to prefer those unusual systems, does not perform significantly better than RS on TREC-{7,8}. An explanation can be found in the indiscriminate mixing of best and worst systems by the DF approach.

We also observed that the problem of underestimating the ranking of the very best systems, decreases considerably for the topic sets of GOV2 compared to TREC Volumes 4+5 and WT10g. While the best performing retrieval system

[1]In previous work, the topic sets TREC-{6,7,8} have mostly been evaluated. No results have been reported yet for topic sets of GOV2.

[2]referred to as $\rho(\mathbf{y}, \mathbf{y}_\mu)$ in [2]

[3]referred to as $\rho(\tilde{\mathbf{y}}, \mathbf{y}_\mu)$ in [2]

(rank 1 in the ground truth) was estimated by the *RS* approach to be ranked at rank 57, 74, 113, 76 and 83 for TREC-{6,7,8,9,10} respectively, the analogous estimated ranking on GOV2 is 30, 32 and 20 for TB-{04,05,06}.

	#sys.	DF	RS	ACScore	ACSim
TREC-6	73	0.600	0.443	0.429	0.425
TREC-7	103	0.486	0.466	0.421	0.417
TREC-8	129	0.395	0.538	0.438	0.467
TREC-9	105	0.527	0.677	0.655	0.639
TREC-10	97	0.621	0.643	0.663	0.649
TB-04	70	0.584	0.708	0.687	0.647
TB-05	58	0.606	0.659	0.547	0.574
TB-06	80	0.513	0.518	0.528	0.458
MEAN		0.542	0.582	0.546	0.535

Table 1: System ranking estimation on the full set of topics. Reported is Kendall's τ. Underlined is the best performing approach per topic set. All correlations reported are significant ($p < 0.005$). Column 2 shows the number of retrieval systems to rank.

In contrast to earlier work, these experiments show that when evaluating a broader set of topic sets of more recent test collections the *RS* method consistently delivers the best overall system rankings.

5.2 Ranking Estimation with Single Topics

In this Section, we show that the ability of system ranking estimation approaches to rank the systems correctly, differs significantly between the topics of a topic set. We set up the following experiment: for each topic, we evaluated the estimated ranking of systems by correlating it against the ground truth ranking that is based on *average precision*. Note, that this is different from the ground truth ranking based on MAP. 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. In Table 2, for each topic set, the minimum (worst topic) and maximum (best topic) τ reached are shown.

The results are similar across all topic sets and system ranking estimation methods: the spread in correlation between the best and worst topic are extremely wide; in the worst case, there is no correlation ($\tau \approx 0$) between the ground truth and the estimated ranking, whereas in the best case the estimated ranking is highly accurate and with few exceptions $\tau > 0.7$. These findings form the motivation for the next section.

5.3 Ranking Estimation with Topic Subsets

We now investigate if we can we improve the accuracy of an estimator when dealing with a subset of topics by for instance removing those topics, the system ranking approach performs most poorly on. Since each of the evaluated topic sets consists of 50 topics we test subsets of cardinality 1 to 50. As it is not feasible to test all possible subsets per cardinality c, for each c we randomly sample 10000 subsets of topics. In total, we test five topic selection strategies, two based on random sampling and three iterative ones:

worst sampled subset: τ of the subset resulting in the lowest correlation

average sampled subset: average τ across all sampled subsets

greedy approach: an iterative strategy; at cardinality c, the 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[4]

median AP: an iterative strategy; at cardinality c the topic is added to the existing subset of $c-1$ topics, that exhibits the highest median average precision across all systems; thus, easy topics are added first

estimation accuracy: an iterative strategy; at cardinality c the 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 (this strategy draws from results of Section 5.2)

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 MAP. In contrast to Section 5.2, now we are 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.

We should stress, that the latter 3 strategies all require knowledge of the true relevance judgments. This experiment was set up, to determine if it is advantageous at all to rely on subsets instead of the full topic set. These strategies were not designed to find a subset of topics automatically.

Exemplary, the results of this analysis are shown for two topic sets in Figure 1[5]. The greedy approach, in particular at subset sizes between 5 and 15, yields significantly higher correlations than the baseline (i.e. the correlation at the full topic set size of 50). After a peak, the more topics are added to the topic set, the lower the correlation. 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 considering the medianAP strategy of first adding easy topics, gains in correlation over the baseline are visible, but they are topic dependent and far less pronounced than the best possible improvement (greedy approach). As hypothesized, adding topics according to the estimation accuracy does indeed lead to improved performance, although the problem remains that without knowing the ground truth it is difficult to estimate what the accuracy of the estimated ranking will be.

In Table 3 the results across all topic sets and approaches are summarized. Across all topic sets and all system rank estimation approaches there exist indeed subsets of topics that would greatly improve the performance of system ranking estimation algorithms. Consider for instance, the results of the *DF* approach on topic set TB-04: with the right topic subset, a rank correlation of $\tau = 0.898$ can be reached, a 54% increase over the performance on the full topic set.

6. CONCLUSIONS

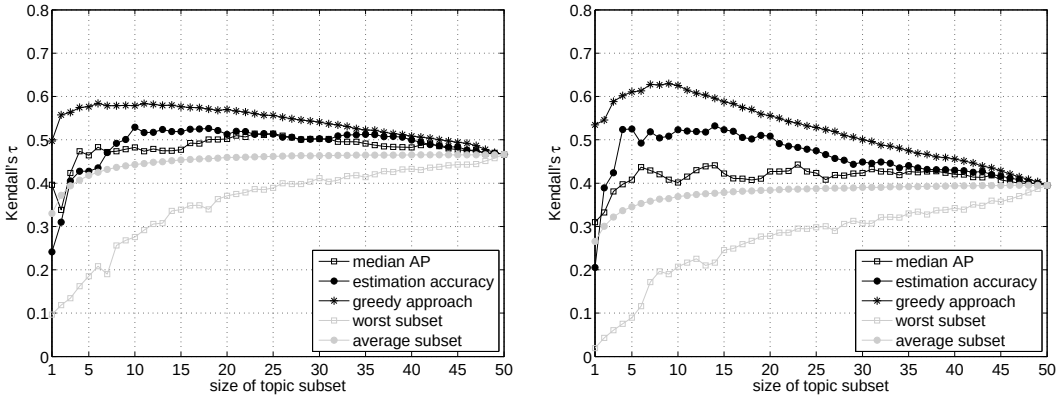
In this work, we have reported the first experiments in topic subset selection for the system ranking estimation task. Contrary to prior work, when widening the number of topic

[4]this approach performs usually as well as or better than the best sampled subset, which is therefore not listed separately

[5]The trends are similar across all topic sets and estimation approaches.

	DF		RS		ACScore		ACSim	
	min. τ	max. τ	min. τ	max. τ	min. τ	max. τ	min. τ	max. τ
TREC-6	0.008	0.849†	−0.106	0.823†	−0.161	0.812†	−0.134	0.777†
TREC-7	−0.061	0.765†	−0.004	0.693†	0.053	0.695†	−0.061	0.687†
TREC-8	0.053	0.792†	0.143	0.731†	0.087	0.741†	0.118	0.735†
TREC-9	−0.234†	0.835†	0.179	0.730†	0.018	0.760†	0.021	0.786†
TREC-10	−0.094	0.688†	0.130	0.821†	0.031	0.722†	0.085	0.769†
TB-04	0.002	0.906†	−0.025	0.882†	−0.161	0.777†	−0.038	0.704†
TB-05	0.040	0.769†	−0.083	0.827†	−0.161	0.717†	−0.052	0.721†
TB-06	−0.070	0.728†	−0.152	0.760†	0.055	0.644†	−0.112	0.701†

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



(a) *RS*, TREC-7 (b) *DF*, TREC-8

Figure 1: Topic subset selection experiments.

	DF			RS			ACScore			ACSim		
	full set	greedy	±%	full set	greedy	±%	full set	greedy	±%	full set	greedy	±%
	τ	τ		τ	τ		τ	τ		τ	τ	
TREC-6	0.600	**0.804**	+34.0%	0.443	0.654	+47.6%	0.429	0.723	+68.5%	0.425	0.721	+69.6%
TREC-7	0.486	**0.762**	+56.8%	0.466	0.584	+25.3%	0.421	0.591	+40.4%	0.417	0.588	+41.0%
TREC-8	0.395	0.630	+59.5%	0.538	**0.648**	+20.4%	0.438	0.606	+38.4%	0.467	0.642	+37.5%
TREC-9	0.527	**0.800**	+51.8%	0.677	0.779	+15.1%	0.655	0.780	+19.1%	0.639	0.786	+23.0%
TREC-10	0.621	**0.761**	+22.5%	0.643	0.734	+14.2%	0.663	0.755	+13.9%	0.649	0.742	+14.3%
TB-04	0.584	**0.898**	+53.8%	0.708	0.846	+19.5%	0.687	0.829	+20.7%	0.647	0.803	+24.1%
TB-05	0.606	0.800	+32.0%	0.659	**0.812**	+23.2%	0.547	0.743	+35.8%	0.574	0.747	+30.1%
TB-06	0.513	0.682	+32.9%	0.518	0.704	+35.9%	0.528	**0.707**	+33.9%	0.458	0.670	+46.3%
MEAN	0.542	**0.767**	+41.5%	0.582	0.720	+23.7%	0.546	0.717	+31.3%	0.534	0.712	+33.3%

Table 3: Summary of topic subset selection experiments. Underlined is the highest correlation of the full set (per topic set) and in bold is the highest correlation of the greedy subset. The columns marked with ± show the percentage of change between τ achieved on the full topic set and the greedy approach. All correlations reported are significant ($p < 0.005$).

sets and TREC corpora, we found the initially proposed random sampling approach [6] to be the most stable and the best performing method. Furthermore, we found that the ability of system ranking estimation approaches to estimate the ranking of systems for each individual topic varies widely within a topic set. Using different topic subset selection strategies, we confirmed the hypothesis that system ranking estimation approaches can be substantially improved, if the "right" topic subsets is used during the estimation. However, this work can only be seen as a first step; for these insights to be useful in practice, automatic methods are required that can identify the best subsets of topics for system ranking estimation. This direction will be explored in future work.

Acknowledgment: We would like to thank Guido Zuccon for his feedback and comments on this work.

7. REFERENCES

[1] J. A. Aslam and R. Savell. On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In *SIGIR '03*, pages 361–362, 2003.

[2] F. Diaz. Performance prediction using spatial autocorrelation. In *SIGIR '07*, pages 583–590, 2007.

[3] J. Guiver, S. Mizzaro, and S. Robertson. A few good topics: Experiments in topic set reduction for retrieval evaluation. *To appear in TOIS*.

[4] S. Mizzaro and S. Robertson. Hits hits trec: exploring ir evaluation results with network analysis. In *SIGIR '07*, pages 479–486, 2007.

[5] R. Nuray and F. Can. Automatic ranking of information retrieval systems using data fusion. *IPM*, 42(3):595 – 614, 2006.

[6] I. Soboroff, C. Nicholas, and P. Cahan. Ranking retrieval systems without relevance judgments. In *SIGIR '01*, pages 66–73, 2001.

[7] A. Spoerri. Using the structure of overlap between search results to rank retrieval systems without relevance judgments. *IPM*, 43(4):1059 – 1070, 2007.

[8] S. Wu and F. Crestani. Methods for ranking retrieval systems without relevance judgments. In *SAC '03*, pages 811–816, 2003.