A few good topics:
Experiments in topic set reduction for retrieval evaluation

JOHN GUIVER
Microsoft Research Cambridge
and
STEFANO MIZZARO
University of Udine
and
STEPHEN ROBERTSON
Microsoft Research Cambridge

We consider the issue of evaluating information retrieval systems on the basis of a limited number of topics. In contrast to statistically-based work on sample sizes, we hypothesise that some topics or topic sets are better than others at predicting true system effectiveness, and that with the right choice of topics, accurate predictions can be obtained from small topics sets. Using a variety of effectiveness metrics and measures of goodness of prediction, a study of a set of TREC and NTCIR results confirms this hypothesis, and provides evidence that the value of a topic set for this purpose does generalise.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms: Measurement, Experimentation

Additional Key Words and Phrases: Search effectiveness, topic selection, evaluation experiments, test corpora

1. INTRODUCTION

In the experimental evaluation of information retrieval (search) systems, we typically take a collection of documents and a set of queries or topics representing information needs. Then we obtain relevance assessments from the originators of the queries or from some substitute judges, on a pooled set of documents. All of this is in essence a well-established methodology, defined in the Cranfield experiments over 40 years ago and developed in TREC and other public research. However, many details of this methodology are open to modification and further development.

Authors' addresses: Stephen Robertson (corresponding author), John Guiver, Microsoft Research, 7 JJ Thomson Avenue, Cambridge CB3 0FB, UK; email {ser,joguiver}@microsoft.com; Stefano Mizzaro, Dept. of Mathematics and Computer Science, University of Udine, Via delle Scienze, 206, 33100 Udine, Italy; email mizzaro@dimi.uniud.it.

One major issue with this methodology is its expense. In particular, the effort required to make relevance assessments is a major bottleneck. Any advance that enables us to get equally good evaluation information with fewer relevance judgements can give us major benefits.

This paper addresses one issue in this general domain: in what ways is it possible to reduce the number of topics needed for evaluation? One might treat this question, as indeed other authors have done, purely as a statistical sampling question: What is a minimal topic set sample size for valid results? In other terms, if one randomly chooses a set of topics, how many topics are needed to obtain statistically reliable results, with a certain confidence? Here, however, we consider another aspect of this issue that, perhaps surprisingly, has been neglected so far: Are some topics (or topic sets) more useful than others for evaluation purposes? Could we choose particular small topic sets which nevertheless are highly informative about the effectiveness of different systems? Can we use such small sets to predict system effectiveness on other topics?

Some evidence presented elsewhere [Mizzaro and Robertson 2007] suggests strongly that individual topics vary greatly in their ability to discriminate between systems; in the present paper, we concentrate on *sets* of topics. However, our approach to this question will not immediately seek a complete solution to the problem of identifying suitable topic sets. Rather we set out to explore the space: to gain some understanding of how different topic sets might give us different amounts of information. As the basis for this exploration, we mine a set of TREC data, representing results on 50 topics and 129 system runs. In addition, we repeat some experiments on a set of data from NTCIR6, consisting of 50 topics and 74 runs, a mixture of mono- and cross-lingual runs.

This paper is structured as follows. Section 2 gives a brief review of some related work. Section 3 gives an overview of our approach to the problem, with some illustrative results. In Section 4 we discuss a number of methodological issues and describe the methods used. Section 5 gives results of a series of experiments on the TREC data, designed to address the methodological issues; in section 5.7 we repeat some of the runs on the NTCIR data, and also use a different metric, based on graded relevance. We conclude in Section 6.

2. REVIEW OF RELATED WORK

There has been concern over many years with how many topics might be needed to enable reliable effectiveness comparisons between systems. Interest in this and related questions is currently going through a major surge; we do not attempt a comprehensive review here, but give an overview, as well as indicating those studies that have contributed directly to the present work.

In 1976 Sparck Jones and van Rijsbergen [1976] claimed that "concerning requests, under 75 are of little use, 250 are usually acceptable, and 1,000 are sometimes needed". Zobel [1998] by contrast, while concentrating on pool depth, did show that results on one set of 25 topics did a reasonable job of predicting relative performance on a different set of 25. Buckley and Voorhees [2000] give evidence that "the number of queries needed for a good experiment is at least 25, and 50 is better" (with differences between different effectiveness metrics). Voorhees and

Buckley [2002], followed by Sanderson and Zobel [2005], give error rate curves for different numbers of topics (similar curves are used below). While studying the statistical power in the context of retrieval evaluation, Webber et al. [2008] conclude that 150 topics are required to distinguish TREC systems in a reliable way.

This work can also lead to conclusions about the choice of effectiveness metrics – some metrics are better than others, in the sense that they require smaller samples of topics for similar accuracy of prediction. Sakai [2007] uses a similar method to Voorhees and Buckley but on metrics which use graded (rather than binary) relevance judgements, and concludes that such measures are stable and sensitive. In another paper [2006], he proposes a different method of evaluating metrics, based on the bootstrap method in statistical significance testing, with some advantages over the Voorhees/Buckley method. Webber et al. [2008] conclude that measures such as Average Precision are better than Precision at rank 10, but go further and argue that AP allows for a better prediction of P@10 on a new set of topics than P@10 itself; thus reporting P@10 on an evaluation set is redundant.

There has also been much recent work (e.g. [Carterette et al. 2006; Yilmaz and Aslam 2006]) on a related efficiency issue: can we reduce the number of relevance judgements per topic and still get reliable results? Methods based on these ideas are currently in use in various TREC tracks.

A more complex series of ANOVA analyses by Banks et al. [1999] reveal that, in addition to a strong system effect, there is an even stronger topic effect. There are also significant interactions between systems and topics, and some weak clustering among topics. All these conclusions are highly relevant as background to the present study. A recent paper [Mizzaro and Robertson 2007] shows through network analysis that some individual topics are much better than others at predicting system effectiveness averaged over all topics. No analysis is reported there of *sets* of topics, but the results on individual topics strongly suggest that similar differences may be expected among sets also.

The object of the present paper is to conduct an exploration of the idea, not addressed by any of the above, that some *sets* of topics are better than others, in the sense of estimating or predicting the general effectiveness of different systems, or at least their relative effectiveness. In the course of this study, we draw upon ideas and methods used in several of the above studies. There is huge scope in investigating this question, and we have only just begun to scratch the surface. We do not at this stage propose any concrete new ways of selecting or devising topic sets for experiments. Nevertheless, we believe that the results demonstrated here provide some insight into these ideas, which may eventually yield new methods.

3. BASIC APPROACH

In this section, we give an overview of our approach to the problem. Here we will give only an outline of the methods used and results obtained, without discussing the various issues involved. The purpose is to present the reader with a general picture of the work, and to avoid getting too deeply into the detail at this stage. Subsequent sections will discuss many details and issues.

Consider Table I, representing a set of TREC (or similar) results. Each row corresponds to a system. (Actually, each row corresponds to a *run*: the same

system, with different parameters, configurations, etc., can lead to several runs.) Each column corresponds to a topic (i.e., request). Therefore, each cell of the matrix $\mathrm{AP}(s_i, t_j)$ measures the performance of system s_i on topic t_j; the classical metric used in TREC is Average Precision (AP). The performance of a system s_i is usually obtained by computing the arithmetic mean of *all* $\mathrm{AP}(s_i, t_j)$ values (one row of the table). This is called Mean Average Precision (MAP).

APs	t_1	$\cdots$	t_n	MAP
s_1	$\mathrm{AP}(s_1, t_1)$	$\cdots$	$\mathrm{AP}(s_1, t_n)$	$\mathrm{MAP}(s_1)$
s_2	$\mathrm{AP}(s_2, t_1)$	$\cdots$	$\mathrm{AP}(s_2, t_n)$	$\mathrm{MAP}(s_2)$
$\vdots$		$\ddots$		$\vdots$
s_m	$\mathrm{AP}(s_m, t_1)$	$\cdots$	$\mathrm{AP}(s_m, t_n)$	$\mathrm{MAP}(s_m)$

Table I. AP and MAP, for n topics and m systems

The data sets actually used in the experiments reported here are discussed below.

The basic method is as follows. We start from a set of n topics ($n = 50$ or 25 in the experiments below). We now consider, for any $c \in \{1, \ldots, n\}$ and for any subset of topics of cardinality c, the corresponding values of MAP for each system calculated on just this subset of topics: that is, we average only a selected set of c of the n columns in Table I. For each such subset, we calculate the correlation of these MAP values with the MAP values for the whole set of topics. This correlation measures how well the subset predicts the performance of different systems on the whole set.

Now for each cardinality c, we select the best subset of topics, that is the one with the highest correlation. We also select the worst, and finally we calculate an average correlation over all subsets of size c. The resulting graphs of correlation against cardinality are shown in Figure 1.

From this figure, we see that the best subset of (for example) size $c = 5$ or 10 is much better at predicting performance on the full 50 topics than a random set of the same size, which in turn is very much better than the worst. To read the figure across instead of down, if our target is 0.95 correlation with the full set, choosing the best subset will allow us to get away with just 6 topics, while if we chose a random subset we would have to go to 22 topics, and if the worst, 41 topics would be required.

We note that these differences are consistent over different topic set sizes and surprisingly large; at first glance, at least, they suggest that some topic subsets are *very much* better than others at predicting effectiveness on the full set. We will need to be a little wary of such differences, because we have also chosen these 'best' subsets from a very large number of candidate subsets; this might be seen as inviting overfitting. However, this matter is discussed and investigated further below.

Thus we have some evidence that a judicial choice of topics will allow us to use much smaller topic sets than random choice, and still have some confidence in the results. However, this result can be criticised in a number of ways, and gives rise to many methodological questions. In the next section, we discuss the various

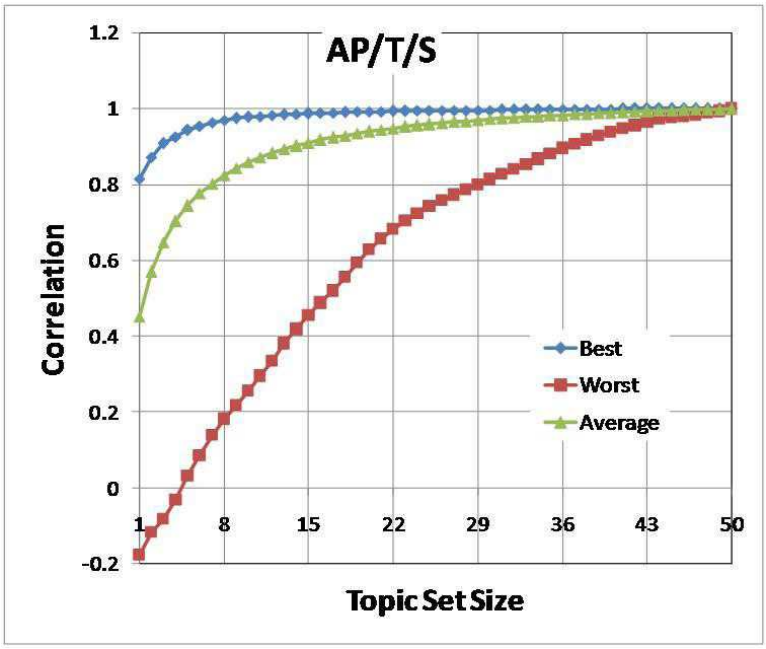

Fig. 1. Maximum, average, and minimum correlation values over cardinalities for AP

methodological issues, before presenting the results in detail. The major issues may be classified as follows:

—Finding the best and worst sets: exhaustive and heuristic search.

—Metrics for system effectiveness (MAP in the example above).

—The measurement of goodness of a subset (correlation in the example above).

—The issues of overfitting and generalisation.

—Computational issues.

In an attempt to reduce potential confusion, we refer below to measures of system effectiveness as 'metrics' and measures of goodness of a subset as 'measures'. We also refer throughout to 'good' or 'bad', or 'best' or 'worst', subsets of topics, meaning those which predict well or poorly effectiveness on some criterion set of topics.

4. DATA AND METHODS

4.1 Datasets

The main dataset used in our experiments is from TREC 8, and consists of 50 topics and 129 systems / runs. Following Voorhees and Buckley [2002], we eliminate from the data the 25% worst-performing runs on the basis of MAP, keeping 96 runs. This will allow us to compare some of our data to [Voorhees and Buckley 2002].

In order to validate (or otherwise) our results, we repeat some of the analyses on a rather different dataset, from NTCIR. This consists of 50 topics and 74 systems

/ runs. As for TREC, we eliminate 25% of the runs, leaving 56. The actual rules for eliminating runs are slightly different, because the table of results by run and topic contains quite a number of zero entries. This appears to be because some runs failed completely on some topics, and no returns were provided. We therefore first removed runs for which more than 2 topics had zero results, and then culled runs with the lowest MAP scores, to a total of 25%.

4.2 Exhaustive and heuristic search

Our basic method is to do an exhaustive search on *all* the possible subsets of topics of cardinality c, for each $c \in \{1, \ldots, n\}$. For each such subset σ and each system s_i, we calculate $\mathrm{MAP}_\sigma(s_i)$, i.e., the MAP for a system s_i computed by averaging the c $\mathrm{AP}(s_i, t_j)$ values of the t_i topics in the σ set only. Then (for each subset) we calculate the correlation of the set of $\mathrm{MAP}_\sigma(s_i)$ with the set of $\mathrm{MAP}(s_i)$, i.e., MAP for system s_i on all the n topics. We choose, out of all subsets of cardinality c, the 'best': the one which gives the highest correlation.

The exhaustive search of the space allows us to find the 'worst' subset as well: the one which gives the lowest correlation; thus, we will know the optimum (how well we can do with $c \leq 50$ topics only), the minimum (how badly we can do if we are extremely unlucky) and the distribution and average (how good we expect to do by random sampling).

Note that the number of subsets of cardinality c of a set of cardinality n is $\binom{n}{c} = \frac{n!}{c!(n-c)!}$. This can be a very large number (for example, $\binom{50}{15} \approx 2 \times 10^{12}$ and $\binom{50}{25} \approx 1 \times 10^{14}$). Thus we do not in fact do a complete search for all $c \in \{1, \ldots, 50\}$.

We could imagine various ways of doing heuristic searches of this space, thus avoiding complete searches. We have used a particular heuristic method, described in Section 4.6 and validated in Section 5, for many of our experiments, including the one whose results are given in Figure 1.

4.3 Effectiveness metrics

As will be very familiar to people, there are many measures (metrics in our current terminology) of retrieval effectiveness (in addition to MAP) which we might use in this context. We consider the following:

—MAP: Mean average precision.

—RPrec: Precision at the rank equal to the number of relevant documents.

—P@10: Precision at rank 10.

—GMAP: Geometric mean average precision.

—NDCG: Normalized discounted cumulative gain.

For GMAP, what we actually use is the equivalent metric $\log \mathrm{AP}$, taking its (arithmetic) mean over topics [Robertson 2006]. We consider NDCG only for the NTCIR data, which has graded relevance judgements. The primary reason for using a variety of metrics is to understand whether and to what extent results depend on the metric chosen.

Note that we could use such metrics either to evaluate a chosen subset, or to choose the optimal subset in the first place. Generally, in the results below, we use the same metric for both purposes (each of the metrics in turn). Another

series of experiments would be needed to decide whether we could use one metric for choosing the subsets to predict a second metric. This would be valuable if we could choose on the basis of P@10, as this is a relatively cheap metric to evaluate; however, this possibility is rendered somewhat less likely by the results of Webber et al. [2008], showing that P@10 is a poor predictor of other measures. This question is left for future work.

4.4 Measuring goodness of a subset

It will be clear that for the purposes of this project, we are taking the full set of results on the $n = 50$ topics / $m = 96$ runs as the gold standard; a subset is good insofar as it reproduces the results on the full set. (Below we also define some experiments with different targets.)

4.4.1 *Correlation.* Given that we are using the MAP effectiveness metric, we would like the $\mathrm{MAP}_\sigma(s_i)$ values from our subset of topics σ to predict the $\mathrm{MAP}(s_i)$ values from the whole set. An obvious way to measure this prediction is to measure the correlation of the two sets of values, as indicated above. However, other methods might be appropriate. Here we consider two more.

4.4.2 *Kendall's τ.* The purpose of measuring $\mathrm{MAP}(s_i)$ is to evaluate systems in a *relative* sense – that is, to help us decide whether system A is better than system B. Thus, it may be argued, we are interested in the rank order of systems as defined by the metric, rather than the values of the metric themselves. Hence we consider as alternative goodness measure the Kendall's τ rank correlation coefficient.

4.4.3 *Error rate.* Decisions about whether one system is better than another are however slightly more complex: we may regard small differences in a metric as not significant by some definition. This question has been addressed in [Sanderson and Zobel 2005; Voorhees and Buckley 2002], where an 'error rate' measure is defined for this purpose. This measure, which actually takes the form of a family of curves on a graph, is defined in Section 5.6.3 below, where we give the results of some similar experiments.

We want to use a similar measure as a goodness criterion in our experiments, but for this purpose we need a single-figure measure for choosing best subsets of topics. This leads us to define a Weighted Average Error Rate as follows. Assume a set of systems or runs $\{s_i, i = 1, \ldots, m\}$. As in the original error rate measure, we take two evaluations of this set, producing two sets of scores according to the chosen metric, $\{X_i\}$ and $\{Y_i\}$. We regard one of these, the Y set, as the ground truth, and evaluate how well the X set predicts this ground truth. A *discordant* pair of runs is one which X ranks in the opposite order from Y; we define an indicator variable for discordance as follows:

$$\delta_{i,j}(X, Y) = \begin{cases} 1 & \text{if } X_i - X_j \text{ and } Y_i - Y_j \text{ are of opposite signs} \\ 0 & \text{otherwise} \end{cases}$$

Again as in the original, we regard this as a function of the difference in performance between the two runs according to ground truth Y. Thus the error rate counts the number of incorrect predictions (discordant pairs), but weights each one by the

ground truth difference:

$$\frac{\sum_{i>j} \delta_{i,j}(X,Y)|Y_i - Y_j|}{\sum_{i>j} |Y_i - Y_j|}. \quad (1)$$

Thus a discordant pair counts more if the true result has a wider margin $|Y_i - Y_j|$ – this is consistent with the intuition behind the original error rate measure. Note that this measure looks quite similar to Kendall's τ, but differs in respect of its dependence on the margin $|Y_i - Y_j|$. In our case, we take the Y set to be the results according to a full set of topics, and the X set as the results from a subset.

Again, we could use a goodness measure either retrospectively, to evaluate a chosen subset, or to choose the optimal subset in the first place, or both. Again we concentrate on experiments which use the same goodness measure for the two purposes; however, we report one indicative experiment in which different measures are used.

4.5 The overfitting issue and generalisation

When we apply the method as discussed so far, to a set S of systems / runs and a set T of topics, some of the ground truth we use to evaluate our subsets is represented in the data we use to choose the subsets in the first place. This presents a danger of overfitting – of choosing some subsets which are purely accidentally correlated with the full set, rather than because of some intrinsic property.

To put this issue another way, suppose we were to construct a relatively sophisticated null hypothesis. This would be based on the notion that in principle no subset is better than any other subset of the same cardinality at predicting full-set performance. However, it would not actually claim that every such subset would have the same correlation with full-set performance – rather, it would anticipate some random variation, resulting in a distribution of values of the goodness measure. Now our combinatorial selection method would find the extreme values of this distribution, over a very large set of candidate subsets. These extremes might then be quite far from the distribution mean.

However, such a null hypothesis would predict that any such result would have no generalisability. The identified extreme subsets would not have any real predictive value beyond the average, when applied for example to held-out data. We attempt to address this question in three ways, essentially holding out data for test according to different schemes.

4.5.1 *Splitting the runs.* First, we conduct the following experiment. We randomly divide the *systems / runs* into two equal disjoint subsets, S_1 and S_2 (48 runs each). For any $c \in \{1, \ldots, 50\}$, we then apply the analysis described above to S_1 to choose the best c topic subset. Then the S_2 systems are evaluated on the chosen topic subset, and we assess the various goodness measures on S_2 only. Specifically, we take as ground truth the set of S_2 results on the full topic set, and evaluate the set of S_2 results on the chosen c topic subset. This should provide a test of generalisation. Note that in these experiment, as in those done with the full set of systems, at cardinality $c = 50$ (all topics), we are guaranteed to reach the ground truth.

There are two issues with this method, namely that (a) the set of TREC runs we

are using includes a number of groups of closely-related runs, essentially generated by perhaps minor variants of the same system, and (b) these same runs were used to generate the pool of documents for relevance assessment for each topic. Thus we might expect some overfitting to happen here as well, insofar as similar runs may occur in both halves. It might be possible to address the first issue by some slightly more complex method of splitting the runs, and perhaps the second by testing on some other set of runs not used in the original pooling. However, in order to provide further evidence on the overfitting question, less subject to these objections, we conduct a second series of experiments splitting the topics.

4.5.2 *Splitting the topics.* This time, we split the *topic set* randomly into two equal disjoint parts (25 topics each). We choose the best c topic subset (for any $c \in \{1, \ldots, 25\}$) from set T_1 – that is, best as evaluated on full T_1 results only. We then evaluate it on set T_2, compared to various possible baselines. Specifically, again, we take as ground truth the set of S results on the T_2 subset, and evaluate the set of S results on the c subset of T_1. Note that in these experiments, using all topics in T_1 does not guarantee that we reach the ground truth goodness of T_2; thus the high-cardinality end of the curve remains below the optimum.

4.5.3 *The Voorhees-Buckley method.* The previously cited work [Sanderson and Zobel 2005; Voorhees and Buckley 2002] takes a view of overfitting / generalisation which is based on topics, like our second approach. They take two disjoint random subsets of topics and consider whether one predicts performance on the other. We present some results using a similar method. We note that this method can be applied only up to a maximum topic subset size of half the total number of topics. Sakai [2006] has developed a bootstrap method which can go all the way up to the full topic set; we have not yet tried this approach.

4.6 Computational issues

As indicated, computation time is an issue for some cardinalities in the exhaustive search scenario. Here we consider a heuristic form of search, which in fact we use extensively.

We observe below that there is some stability in best sets. That is, if we consider the best set at cardinality $c = 8$ and the best set at $c = 9$, it is likely that they overlap considerably: it could be that the latter consists of the former together with one additional topic. Thus a heuristic is to search recursively: having identified the best set for cardinality c, to seek the best for cardinality $c + 1$ among sets which differ from the best c set by not more than 3 topics (the number 3 was chosen primarily because 4 is intractable).

We found it necessary to use such a method for cardinalities $10 < c < 41$ in the 50 topic set. For example, when using the Kendall's τ, searching exhaustively takes around 7 days for $c = 11$, and around 20 days for $c = 12$, even with efficient $O(n \log n)$ calculation of τ using Knight's algorithm [Boldi et al. 2005]. Exhaustive search for correlation runs is much faster due to the simpler calculation, and wider scope for optimisation of the algorithm; however, even there, computation becomes a real issue beyond $c = 15$.

Below we present some evidence that the heuristic works well – even if it does not find the absolute best topic set, it will find something very close to the best.

The method is therefore suggested as a usable general heuristic for this problem, and has been used (except where otherwise specified) throughout the experiments below.

Some alternative heuristics could be considered. The simplest might be just to sample (say) 1000 different sets of c topics and take the best from the sample; this would also be likely to give results close to the optimum, and would avoid the dependence on the chosen $c-1$ set. A slightly more sophisticated method would combine both these heuristics: start by sampling, choose the best in sample, and then seek to improve that by considering small variations, as in the above heuristic. Neither of these ideas has been tried yet.

5. RESULTS

All the results that follow are based on the TREC dataset, until Section 5.7, where we switch to the NTCIR dataset.

5.1 Validation of the heuristic

The heuristic method defined in the previous section for discovering best or worst subsets of different cardinality requires justification. We conducted both heuristic and exhaustive searches for best and worst sets on each of the two subsets of topics T_1 and T_2 for all cardinalities, all four effectiveness metrics, and all three goodness measures; and on the full set T for cardinalities $1 \leq c \leq 10$ and $41 \leq c \leq 50$, for AP only. For each combination from best/worst, topic set, metric, measure we observe the average and maximum score difference between the exhaustive and the heuristic choice, over all cardinalities, as a percentage of the score range.

In the case of average precision, the largest differences for best sets on T_1 are seen with Kendall's τ: here the maximum score difference is 1.19% and the average is 0.077%. Mostly the heuristic best is within the best 10 sets identified exhaustively; occasionally it drops out of the top 10. The heuristic and exhaustive worst sets are usually even closer – for some combinations of measure and topic set no differences at all were discovered. Results for GMAP and RPrec are broadly similar; Precision@10 results appear slightly less stable.

We regard these results as confirming that the heuristic is good enough for our purposes. Some additional comments on the method are given in Sections 5.4 and 5.5 below.

5.2 Linear correlation analysis

In these experiments, we continue to use the simple linear correlation analysis, on all 50 topics and 96 systems, but apply the method with the four different effectiveness metrics. The results are shown in Figure 2 (the first graph reproduces Figure 1). The pattern is very similar for the four metrics: best subsets are better than random subsets which are very much better than worst subsets, especially for low cardinality.

We conclude that the effects we observe are essentially independent of the effectiveness metric used. We note, however, that we have not yet investigated whether the optimum set for one metric is also good for the other metrics.

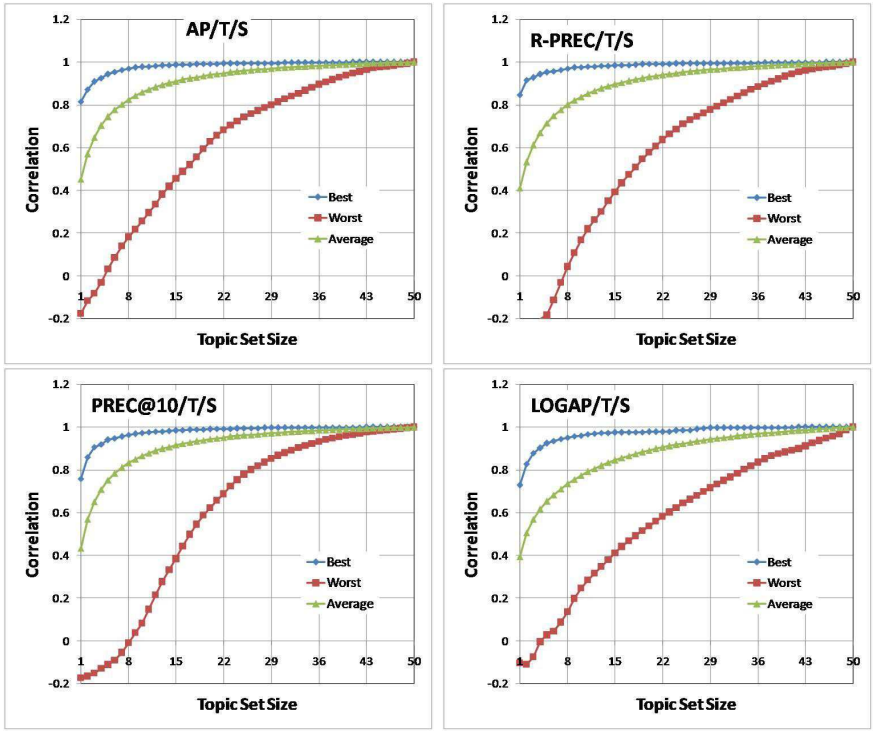

Fig. 2. Correlation: maximum, average, and minimum goodness measure values over cardinalities, for the four effectiveness metrics, Average Precision, RPrec, Precision@10 and GMAP

5.3 Optimising other goodness measures

As discussed above, we may argue that the linear correlation measure is not the best way to decide if a subset of topics predicts some ground truth well. Here we try the same analysis with two alternative measures, Kendall's τ rank correlation and a measure based on Error rate [Voorhees and Buckley 2002].

Kendall's τ is straightforward; we simply substitute τ for the linear correlation used in the previous analysis. The corresponding graphs are shown in Figure 3. The pattern is very similar indeed to the pattern seen for correlation.

For the error rate, the present purpose requires a single-figure measure that can be used to choose a subset unambiguously. We use the Weighted Average Error Rate defined above. Results are shown in Figure 4. The measure is the reverse of the correlation measures (lower is better), but apart from this, the pattern is again very similar.

We conclude that the effects we observe are essentially independent of the measure of goodness used. We note again, however, that we have not yet investigated whether the optimum set for one measure is also good for the other measures.

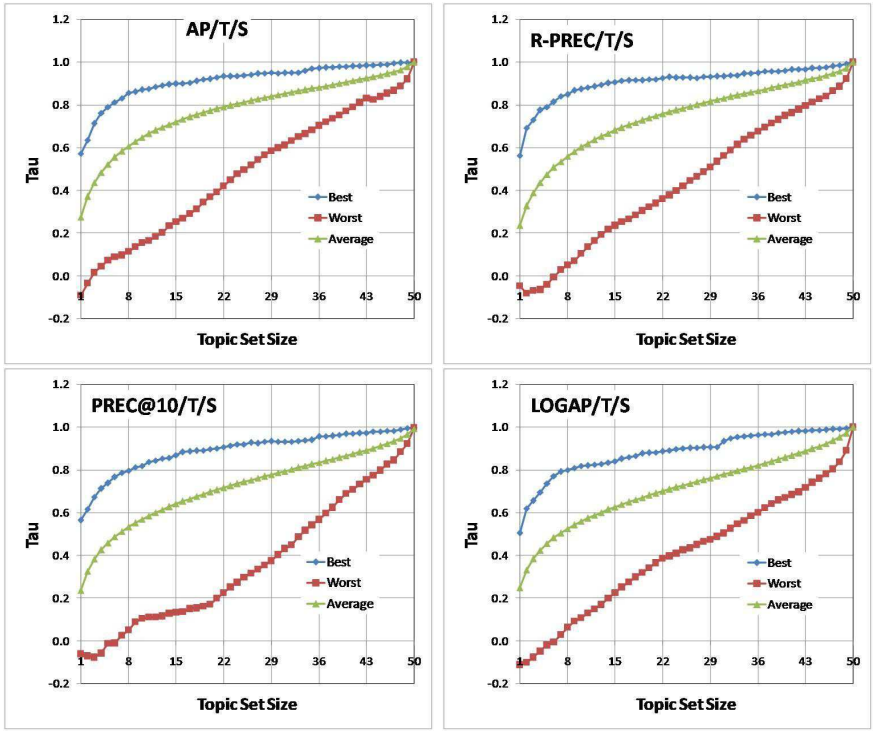

Fig. 3. Kendall's τ: as Figure 2

5.4 Subset analysis

In these experiments, we further investigate the question relating to the heuristic search method: How much difference is there between the best set for one cardinality and that for the next? Using the exhaustive searches on the T_1 subset of topics, and the Average Precision metric and the correlation measure, we show a pixel-map (Figure 5) of the occurrences of topics in best/worst subsets for different cardinalities. Each column represents a topic and each row a cardinality. The pattern is very similar for the other metrics and measures.

It appears that on the whole, topics are good or bad – that is, a topic will tend to appear in the good sets or in the bad sets, but not both. However, there are rather different patterns between best and worst: once a topic is in a worst set, it tends to stay there, but there is considerable variation in the best sets. At one extreme, the best at $c = 9$ and the best at $c = 10$ are completely disjoint. It follows that the exhaustive and heuristic searches must diverge at this point – in fact the heuristic best 10 is the second best set of 10 on the exhaustive list (with only a small drop in correlation). The following section gives us further insight into these comparisons.

To summarise this data, for each combination of metric and measure, we measure the stability of the best (respectively worst) sets as follows. For each topic in a set of a given size, a counter is incremented if the topic is in the set next size up. For the exhaustive search, this counter has a minimum and maximum possible value,

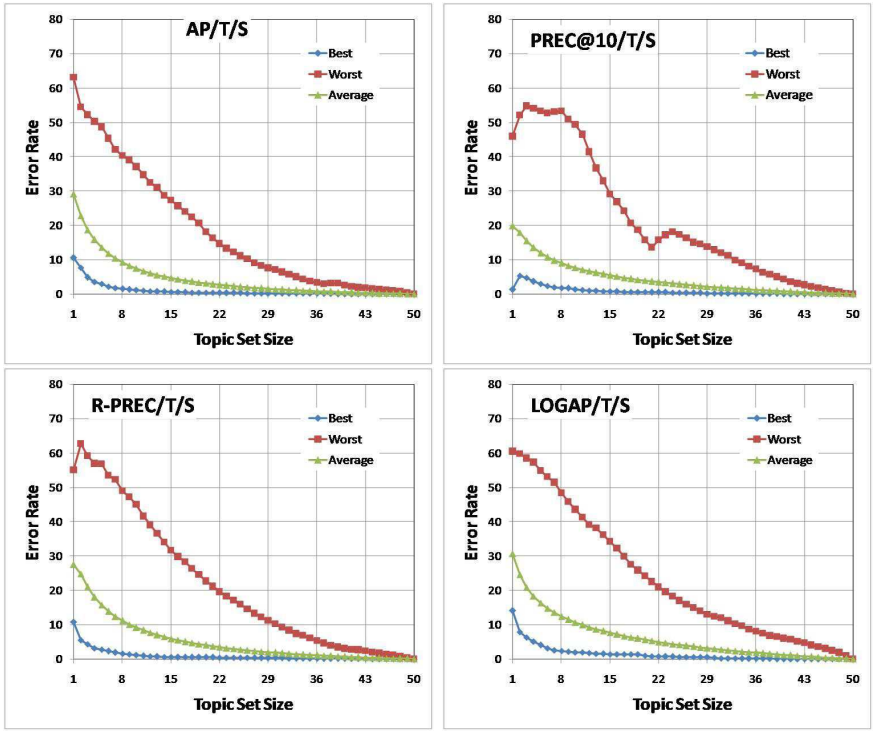

Fig. 4. Error Rate: as Figure 2

where the minimum is greater than 0 due to the fact that as topic sets get larger there is some inevitable overlap from one topic set to the next. The stability value (expressed as a percentage) is given by the ratio:

$$\frac{counter\ \text{actual} - counter\ \text{min}}{counter\ \text{max} - counter\ \text{min}}.$$

For worst subsets the average stability for Average Precision across the three goodness measures is 93%. Best subsets are somewhat less consistent, with an average of 86%.

5.5 Neighborhood analysis

We now seek to compare the best subsets with the not-quite-best, and the worst with the not-quite-worst. We select the ten best (respectively the 10 worst) subsets for cardinality 12, and represent them in similar pixel maps in Figure 6. We observe again, very clearly, that the worst sets show greater consistency than the best.

We hypothesise the following explanation of the results of this and the previous section. While there are some topics that are just plain bad, the quality of being a good topic for this purpose is somewhat more subtle. In particular, topics can have some complementarity. A good *set* is not just a set of topics that are good individually, but topics that complement each other in predicting or explaining variations between systems. This is consistent with the observation in [Banks et al. 1999] that

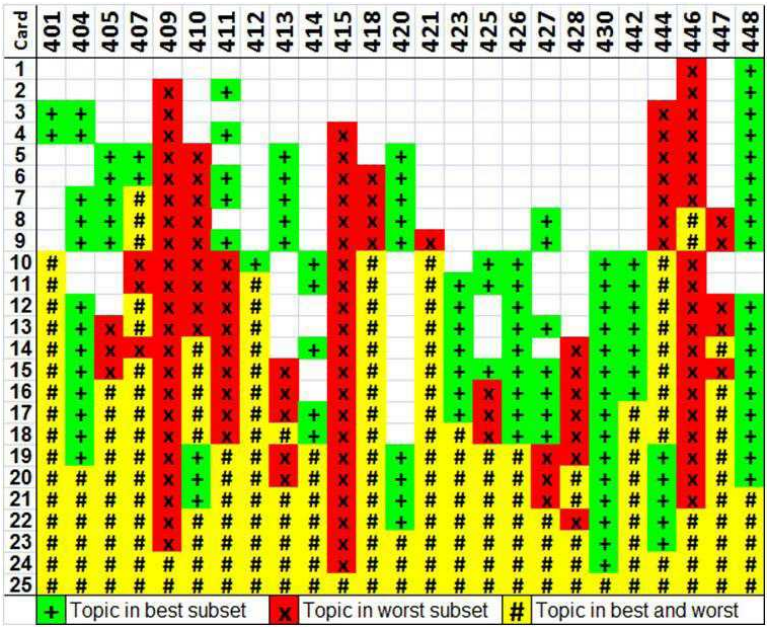

Fig. 5. Pixel map of best/worst sets overlap, average precision / correlation / T_1/S. Columns are topics, rows are cardinalities. For example, for cardinality 2, the worst topic set is {409,446}, and the best is {411,448}

topics cluster weakly: we might expect that, in order to get good prediction, we need a good range of different topics covering multiple clusters. It is also consistent with our observation that some apparently bad topics are also included in good sets: for example, in Figure 6, the rightmost topic (450) is consistently in the bad sets but also appears in some good sets.

Finally, and perhaps surprisingly, the variety implied by this hypothesis is consistent with the success of the heuristic method. While the absolute best set may vary significantly between cardinalities, there is likely to be another somewhat different set which is almost as good, and which is reachable via the heuristic. This is despite the fact that the number of potential subsets is astronomical, and the number explored via the heuristic is tiny by comparison.

5.6 Generalisation experiments

All the preceding analysis has been based on choosing a subset of topics and judging how well it predicts the results from the full set, for a fixed set of systems / runs. We expect that this process results in some overfitting, because the 'ground truth' against which we judge the selection includes the data used to make the selection. In this section, we seek in different ways to separate the two. In a worst case, we might find that the selection is no better than random when judged on an independent, held-out set of ground truth data.

5.6.1 *Hold out systems.* Our first experiment involves holding out some of the systems. As discussed above, we make a random split of the set of systems / runs

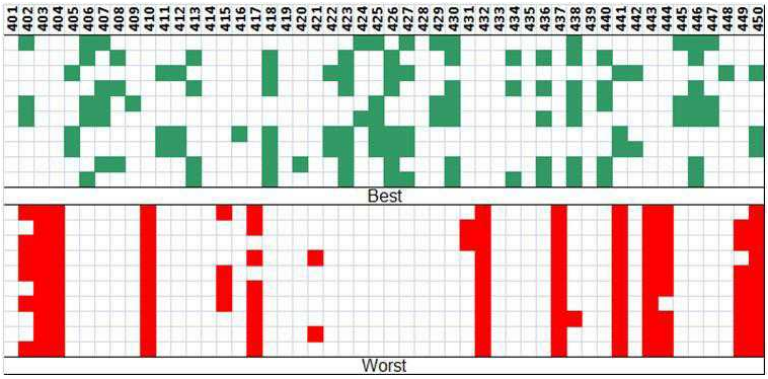

Fig. 6. Best/worst 10 subsets cardinality 12 overlap, average precision / correlation / T/S. For example, topic 402 occurs in 3 of the ten best cardinality 12 subsets, but also in 6 of the 10 worst ones. 403 and 404 occur in all 10 worst sets only.

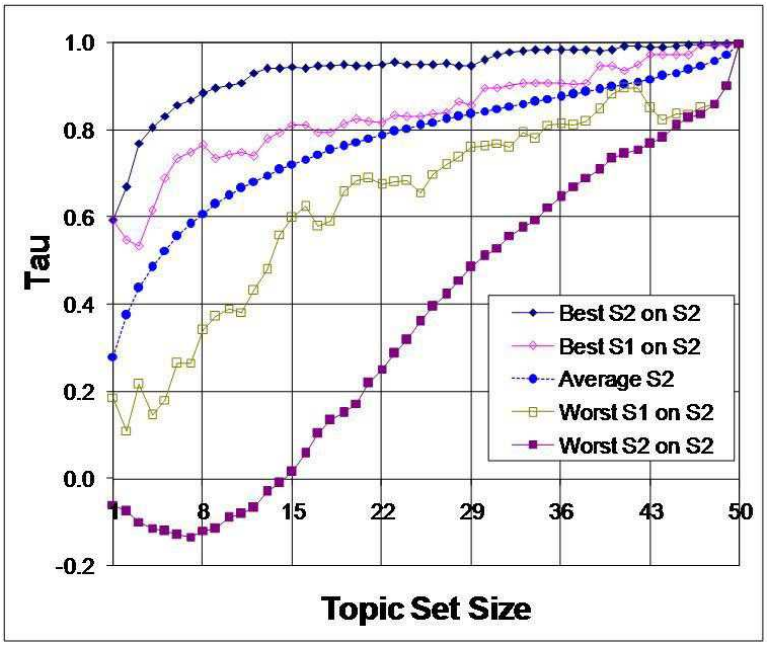

Fig. 7. Maximum, average, and minimum Kendall τ values over cardinalities for AP, all topics, subsets chosen using S_1 or S_2, evaluation on S_2

into two equal subsets; we now choose the best/worst cardinality n subsets of topics according to S_1, and test it on S_2 (compared to the best/average/worst subsets according to S_2). The results for average precision and Kendall τ, all tested on S_2, are shown in Figure 7; results for other effectiveness metrics or goodness measures are very similar.

We see that despite a significant amount of overfitting when the best sets are

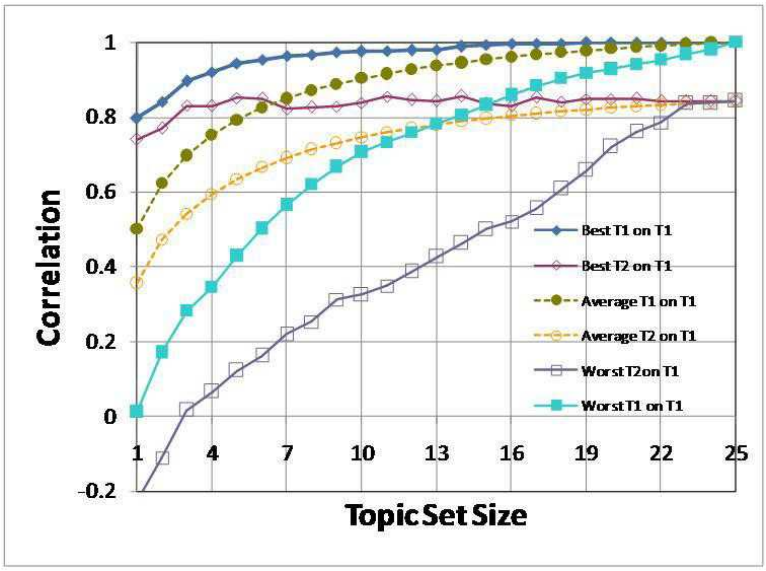

Fig. 8. Best, average and worst, average precision / correlation / T_2/S, evaluated on random T_1 subsets

chosen on their S_2 results, there is still a substantial difference between the best and worst subsets when chosen on their S_1 results. In almost all cases, the best S_1 sets also outperform average subsets. This provides good confirmation that the selected subsets of topics really do have some generalisation properties. Once we have identified a good topic set for a set of systems / runs, this set is now much more effective than a random set of similar size for evaluating further systems / runs.

A possible limitation of this method of holding out systems (actually runs) has already been noted in section 4.5: the set of TREC results used contains some groups of related runs, essentially distinct runs based on perhaps minor variants of a specific system. We did not control for this possibility when creating the random split of systems, so that some overfitting may remain as a result of a group of related runs being divided between the two halves. However, we now show the results of a different hold-out method, not subject to this objection.

5.6.2 *Hold out topics.* In this experiment, we split the topics instead of the systems – we now optimise the subsets of T_2 and test whether they predict system effectiveness on the T_1 topics, all using the entire S set of systems / runs. This situation is slightly different, in that the entire T_2 set (the right-hand end of the curve) will no longer predict the T_1 results perfectly. To put it another way, overfitting remains a strong and unavoidable factor throughout the curve. At the right hand end, even the best T_2 cannot be expected to do as well as random T_1. Figure 8 shows the results for average precision and correlation; the other combinations are very similar.

Again, despite a significant amount of overfitting by the T_1 subsets, there remains a substantial difference between best and worst T_2 subsets. The best T_2 subsets

always do better than average T_2 subsets, although the difference is small at higher cardinalities. For the lowest cardinalities (up to 6), the best T_2 subsets even do better than average T_1 subsets. Again, we have evidence of generalisation. Once we have identified a good topic set for a set of systems / runs, this set is now more effective than a random set of similar size for predicting the performance of these same systems / runs on other topics. For example, it takes only the single best topic from T_2 to achieve a correlation above 75%, while with a random set, we would have to go to cardinality 9.

5.6.3 *Voorhees-Buckley experiments.* The experiments designed by Voorhees and Buckley [2002] work as follows. For each cardinality $1 \leq c \leq 25$, two random non-overlapping sets of c topics are chosen, with one being taken as ground truth. A set of bins (20 bins of width 0.01, going from 0.0 to 0.2) is created; these bins are keyed, for a given pair of systems, by the absolute difference between the effectiveness metric (according to the ground truth) of the two systems. Two counters are initialized to 0; the first is incremented each time a pair of systems belongs to the bin, and the second counter is incremented only if, in addition, the non-ground truth ranking of the two systems is discordant with the ground-truth ranking. This is repeated over all system pairs for many different pairs (10,000) of random non-overlapping sets. At the end, the ratio of the second counter to the first counter, expressed as a percentage, gives an error rate for each bin for the given cardinality. This is now repeated over all cardinalities ($1 \leq c \leq 25$), to give percentage error rates for each bin and cardinality. We do the same experiment, but based on the T_1/T_2 split: random sets from T_2 are taken as ground truth, and the best (according to our error rate measure) or random sets from T_1 are evaluated. A pseudocode version of this procedure is given as Algorithm 1; this shows the distinction between the original Voorhees & Buckley method and the version used here. These error rates are plotted as curves where, for each bin, percentage error rates are plotted against cardinality.

We show in Figure 9 the family of curves for best T_1; the basic pattern is the same as for random T_1 and the same as reported in [Voorhees and Buckley 2002; Sanderson and Zobel 2005]. Comparing best/random/worst quantitatively is hard on such a graph, so we select three bins and plot the two curves for these three bins only, plus the average (Figure 10). Note that each set must share the same right-hand end, because there is only one set of 25 topics. At every point the 'best' curve is a little better (lower) than, or coincides with, the corresponding 'random' curve. This provides additional evidence that selection of best topics really does give us some gain over random sets, in being able to achieve good evaluations with smaller sets of topics. The 'worst' curves show some discrepancies – at high cardinalities they are sometimes slightly better than either 'random' or 'best'. However, in the low-cardinality range where the differences are much more marked, the order is consistent and indeed the 'worst' curves are much worse.

5.6.4 *Comparison across goodness measures.* The number of experiments that could be done in which we select according to one goodness measure and then evaluate according to another is enormous (given also the possible choices of effectiveness metric). In general this is left for future work. However, we show one

Algorithm 1 Voorhees-Buckley experiment

Require: 20 bins k, defined by the interval $(0.01(k-1), 0.01k]$. A system pair will be assigned to the bin corresponding to the absolute difference in ground truth effectiveness score between the two systems.
Require: Two counters per bin, all(k) and disc(k)
  for $c = 1$ to 25 **do**
    for $k = 1$ to 20 **do**
      Initialise all(k) and disc(k) to zero
    end for
    for $r = 1$ to 10000 **do**
      if original Vorhees-Buckley then **then**
        Pick a random subset $\mathcal{X}$ of c topics
        Pick a second random subset $\mathcal{Y}$ of c topics, not overlapping with $\mathcal{X}$
      else if evaluate on random T_1 **then**
        Pick a random subset $\mathcal{X}$ from T_1 of c topics
        Pick a random subset $\mathcal{Y}$ from T_2 of c topics
      else if evaluate on best T_1 **then**
        Pick the best subset $\mathcal{X}$ from T_1 of c topics
        Pick a random subset $\mathcal{Y}$ from T_2 of c topics
      end if
      for $i = 1$ to m **do**
        Evaluate system/run i on $\mathcal{X}$, giving X_i
        Evaluate system/run i on $\mathcal{Y}$, giving Y_i
      end for
      for all i, j such that $1 \leq i < j \leq m$ **do**
        Select bin k according to $|Y_i - Y_j|$
        Increment all(k)
        if $\delta_{i,j}(X, Y) = 1$ (i.e. if the two are discordant) **then**
          Increment disc(k)
        end if
      end for
    end for
    for $k = 1$ to 20 **do**
      Error rate $e_{(c,k)} = \frac{\text{disc}(k)}{\text{all}(k)}$
    end for
  end for

such comparison here, as an indication only. We repeat the previous experiment, but choose the best T_1 subsets according to the correlation criterion rather than the error rate. The results are shown in Figure 11. We observe that the differences between best, random and worst are very similar to those observed above. This result provides at least preliminary evidence that the choice of goodness measure may not be too critical: that a set of topics that is good according to one measure may also be good according to the others.

5.7 NTCIR experiments

In these experiments we repeat some of the previous experiments on the second data set, from NTCIR, and also using a different metric, NDCG.

Figures 12 and 13 contain the results for the NTCIR data, on AP (left) and NDCG (right): in Figure 12, the first row corresponds to Figure 1, the second to Figure 8, and the third to Figure 9. In Figure 13, the first row contains pixel maps corresponding to Figure 5, and the second to Figure 6. The last row contains S_1/S_2

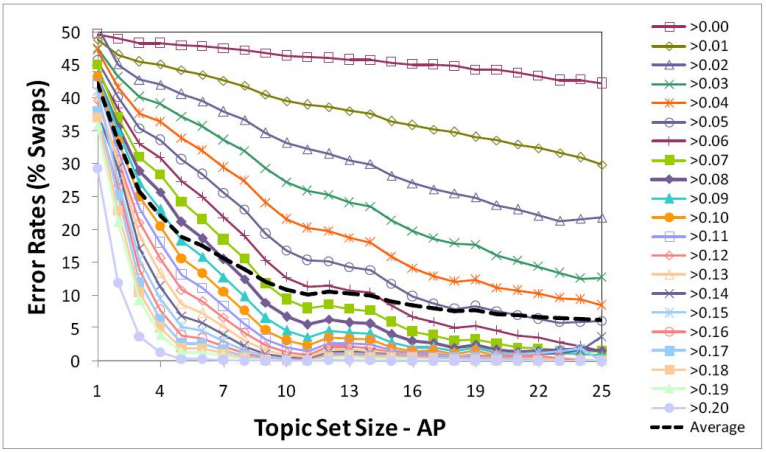

Fig. 9. Best average precision / error rate / T_1/S, evaluated on random T_2 subsets

generalization results corresponding to Figure 7.

In general, these results are similar to those observed using TREC data. Exactly the same comments about the consistency of the best and worst sets respectively apply to NTCIR. The results do however appear to be somewhat noisier, particularly the AP S_1/S_2 results in the last row of Figure 13. This is not really surprising, for two reasons: firstly, we have fewer runs in this set; and secondly, they represent a mixture of mono- and cross-lingual runs. Indeed it is quite surprising that we get any generalization at all – one might expect that the usefulness of a topic set in predicting effectiveness would be very greatly affected by language and translation issues. To put it another way, we might guess that an ANOVA (along the lines of Banks et al. [1999]) on NTCIR data would show a significant topic-language interaction (quite apart from possible system-language and 3-way interactions). However, despite this, there still seems to be some signal present, particularly in the NDCG results.

6. DISCUSSION

6.1 Conclusions

This paper started from the hypothesis that some topics or topic sets are better than others for evaluating systems, in terms of their ability to predict absolute or relative system performance in other data sets. We have shown that given a larger set of topics, it is indeed possible to discover subsets that do have better predictive power than random subsets. We have provided extensive evidence that this is not just a random effect: that good topic sets continue to outpredict random sets of similar size on held-out data (either unseen systems or unseen topics).

This result may be compared to the more statistically-based work in [Voorhees and Buckley 2002; Sanderson and Zobel 2005] and elsewhere. Our result does depend, as the cited work does not, on the availability of a larger set of evaluated topics to select from, together with a set of system runs. This issue is discussed further below.

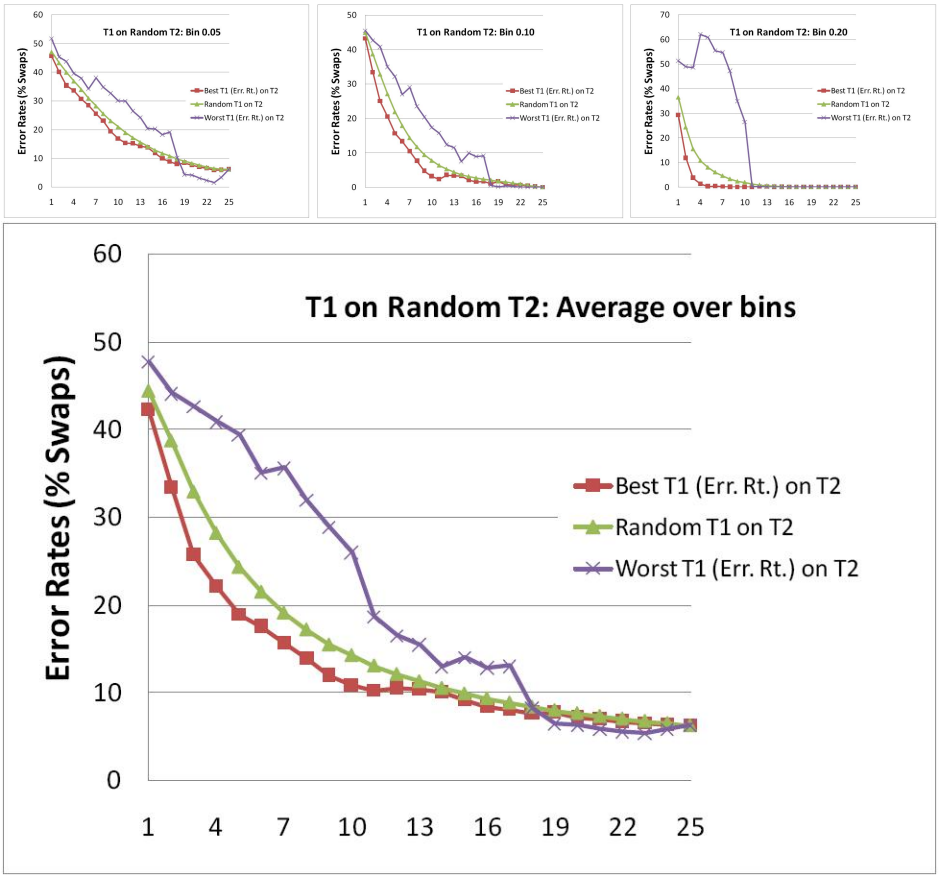

Fig. 10. Best, random and worst, average precision / error rate / T_1/S, evaluated on random T_2 subsets (three bins only plus average)

We may summarise our more detailed results by saying that not all topics are equally informative about systems. However, this statement requires qualification. While some topics are simply bad for this purpose, in some sense the important thing is the *set* rather than the individual topics – it appears that different topics provide complementary evidence about systems.

We have provided some different selection criteria, but no evidence has emerged as to which criterion might be best. All our work so far suggests that different choices of goodness measure and/or effectiveness metric provide comparable results (with the possible exception of Precision@10, which seems somewhat less stable than the other metrics). However, we have only scratched the surface of this space.

6.2 Implications

As indicated, a direct application of the methods used in this paper requires a large set of evaluated topics and associated system runs, and a somewhat complex combinatorial analysis of results. This is not intended as a realistic scenario – rather,

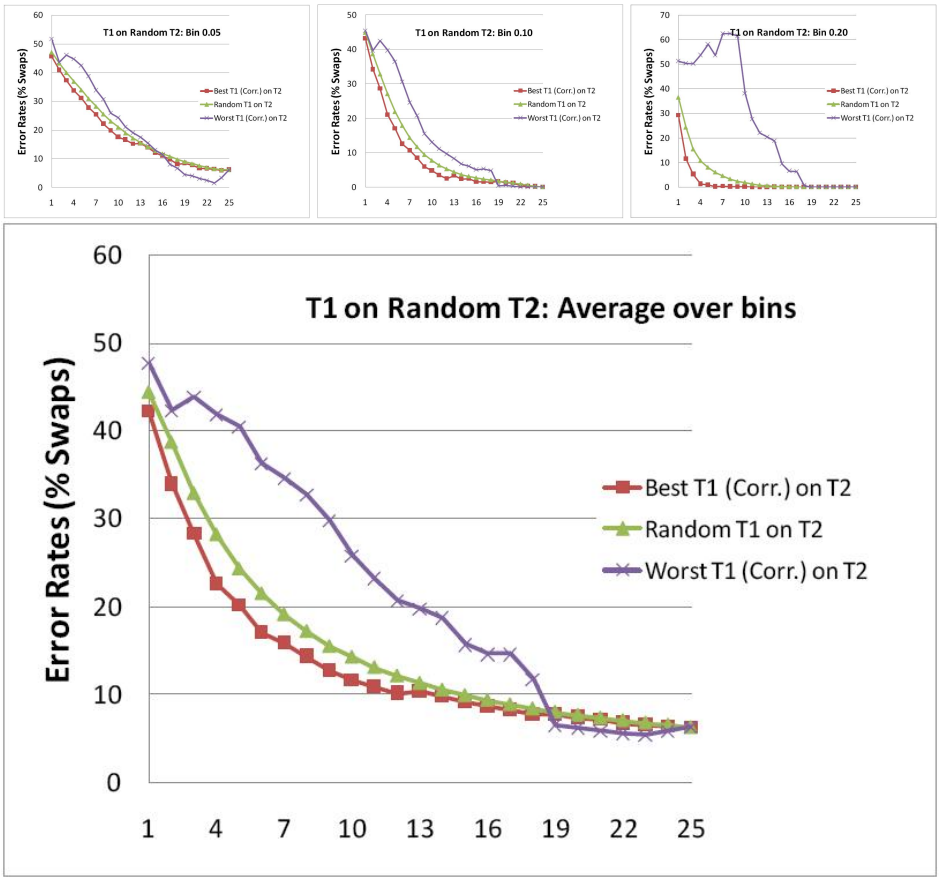

Fig. 11. Comparison across goodness measures: maximum, average, and minimum Error Rate values over cardinalities, subsets chosen on the AP criterion from T_2 or T_1, evaluation on T_1 and S

we have sought evidence that some form of topic selection (other than random) might be valuable. Having provided quite strong evidence for this statement, we now need to investigate possible selection methods and scenarios.

Despite the above, the methods developed in this paper could perhaps be used in the following scenario. An organised effort such as TREC or NTCIR could be conducted in the usual fashion, with the usual selection of topics. Then a subset of topics could be chosen as above, and issued as an 'official' condensed set of topics. Groups could use the condensed set for a wide range of evaluations / optimisations, and verify results on the full set.

The condensed collection would be particularly useful for subsequent experiments requiring manual effort on a per-topic basis (such as interactive experiments, or manual query expansion, etc.). However, for the standard automatic benchmark evaluation, the gain from such condensation may be relatively small, for a significant effort; we therefore discuss here both the effort and (in outline) possible scenarios.

Considering the combinatorial problem, it would be appropriate to investigate

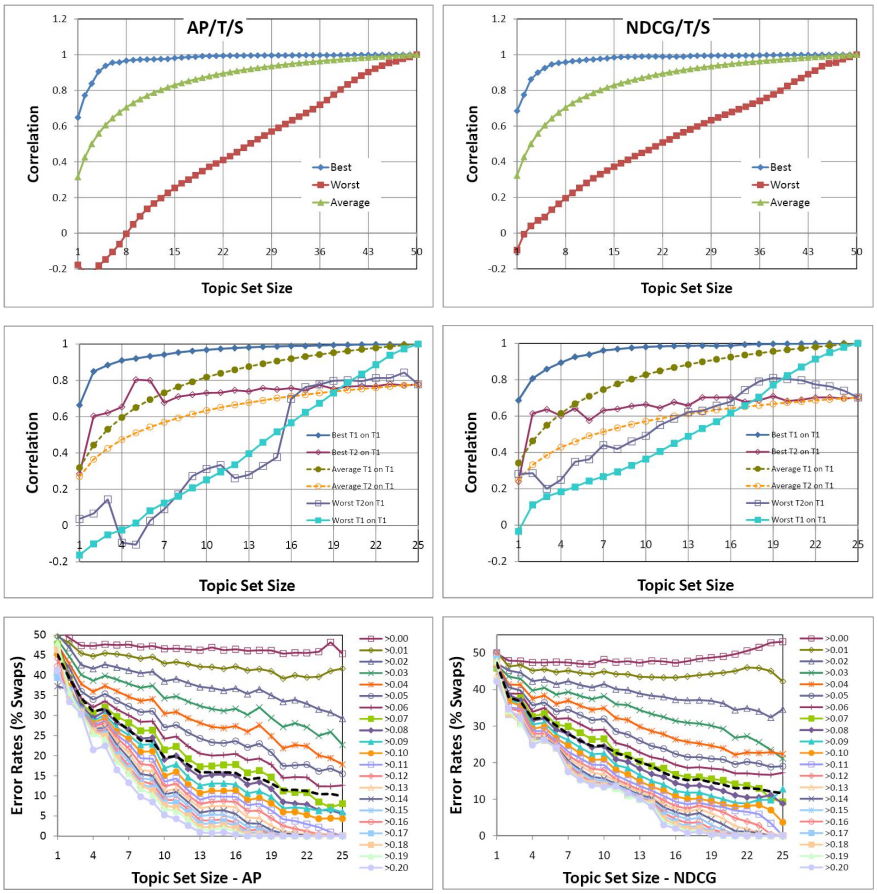

Fig. 12. NTCIR results, AP on left and NDCG on right. First row: correlation goodness measure based on T and S. Second row: correlation goodness measure based on T_2 and S, evaluated on random T_1 subsets. Third row: error rates, based on T_1 and S, evaluated on random T_2 subsets.

per-topic selection criteria, such as for example a measure of how well each single topic predicts average effectiveness. There are several possible ways to measure this – a measure of topic 'hubness' is proposed in [Mizzaro and Robertson 2007], for example. The inference noted earlier concerning the complementarity of topics in the good sets might work against such a method; however, the observation concerning the variety of good topic sets might work in its favour, just as it appears to help the heuristic method.

An alternative, given the relative consistency of the sets of *bad* topics suggested by our results, is to select topics for exclusion rather than for inclusion. It may be easier to make reliable rejection decisions than selection decisions.

The issue of discovering other variables which might be predictive of whether topics or topic sets are good in the present sense is a major one. We do not at this stage, for example, know whether there is any relation between goodness for

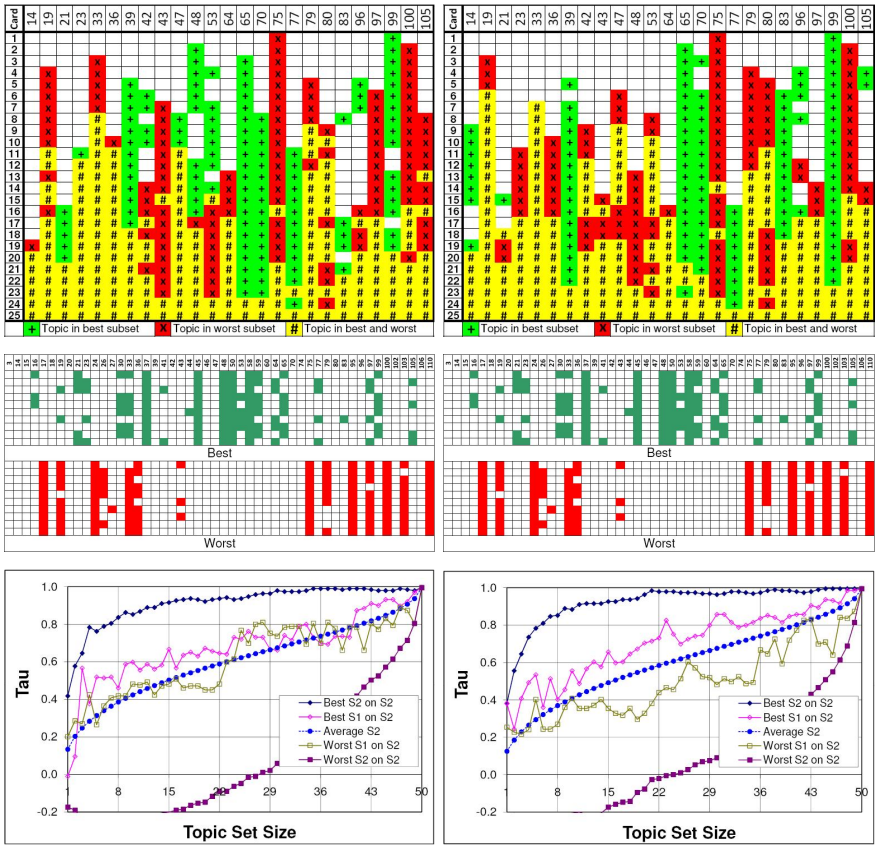

Fig. 13. NTCIR results continued, AP on left and NDCG on right. First row: best and worst sets for different cardinalities, based on T_1 and S. Second row: 10 best and worst sets for cardinality 12, based on T and S. Third row: S_1/S_2 generalization.

the present purpose and topic 'hardness', as investigated for example in the TREC Robust track. We have not yet attempted to replicate the methods proposed here on Robust Track data, as this would have introduced a new dimension to an already complex picture, but that is clearly an investigation worth pursuing.

But the more serious issue concerns the availability of fully-evaluated results in the first place. We can envisage scenarios in which topic selection happens at various different stages. Firstly, there may be intrinsic topic features which are predictive of goodness or badness, or are predictive of some kind of complementarity that might contribute to a good set, and which are accessible before any searching takes place. Secondly we might look at some set or sets of search results (one or multiple systems) prior to relevance assessment. Thirdly, we might seek information during the process of relevance assessment, which might lead to a decision to focus assessment resources on certain topics. These are all methods which have been proposed or used for other search-related decisions, for example, choosing between

ranking algorithms that suit easy or hard topics or topics of different types, or distributing limited assessment effort to optimise average precision estimates. We note in particular recent work on reducing the number of relevance assessments required [Carterette et al. 2006; Yilmaz and Aslam 2006], which should certainly be tied in with the present work.

At this stage we do not know whether any such method could give us any of the benefits of topic set selection that we have seen in this paper. In this sense, the results shown here provide an optimum for us to aim at. But our results do suggest strongly that it is worth investigating such methods.

Acknowledgements

We are very grateful to the anonymous referees for a set of thoughtful and valuable comments, and to Noriko Kando and Tetsuya Sakai for making available the NTCIR data used.

REFERENCES

Banks, D., Over, P., and Zhang, N.-F. 1999. Blind men and elephants: Six approaches to TREC data. *Information Retrieval 1,* 1-2, 7–34.

Boldi, P., Santini, M., and Vigna, S. 2005. Paradoxical effects in PageRank incremental computations. *Internet Mathematics 2,* 3, 387–404.

Buckley, C. and Voorhees, E. 2000. Evaluating evaluation measure stability. In *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, N. J. Belkin, P. Ingwersen, and M.-K. Leong, Eds. ACM Press, New York, 33–40.

Carterette, B., Allan, J., and Sitaraman, R. 2006. Minimal test collections for retrieval evaluation. See Efthimiadis et al. [2006], 268–275.

Efthimiadis, E. N., Dumais, S. T., Hawking, D., and Järvelin, K., Eds. 2006. *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York.

Mizzaro, S. and Robertson, S. 2007. HITS hits TREC — exploring IR evaluation results with network analysis. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, C. L. A. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. P. de Vries, Eds. ACM Press, New York, 479–486.

Robertson, S. 2006. On GMAP – and other transformations. See Yu et al. [2006], 78–83.

Sakai, T. 2006. Evaluating evaluation metrics based on the bootstrap. See Efthimiadis et al. [2006], 525–532.

Sakai, T. 2007. On the reliability of information retrieval metrics based on graded relevance. *Information Processing and Management 43*, 531–548.

Sanderson, M. and Zobel, J. 2005. Information retrieval system evaluation: effort, sensitivity, and reliability. In *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, and N. Ziviani, Eds. ACM Press, New York, 162–169.

Sparck Jones, K. and van Rijsbergen, C. J. 1976. Information retrieval test collections. *Journal of Documentation 32*, 59–75.

Voorhees, E. and Buckley, C. 2002. The effect of topic set size on retrieval experiment error. In *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, M. Beaulieu, R. Baeza-Yates, S. H. Myaeng, and K. Jarvelin, Eds. ACM Press, New York, 316–323.

Webber, W., Moffat, A., and Zobel, J. 2008. Statistical power in retrieval experimentation. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*. ACM, New York, NY, USA, 571–580.

Webber, W., Moffat, A., Zobel, J., and Sakai, T. 2008. Precision-at-ten considered redundant. In *SIGIR 2008: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, S.-H. Myaeng, D. W. Oard, F. Sebastiani, T. Chua, and M.-K. Leong, Eds. ACM Press, New York, 695–696.

Yilmaz, E. and Aslam, J. A. 2006. Estimating average precision with incomplete and imperfect judgements. See Yu et al. [2006], 102–111.

Yu, P. S., Tsotras, V. J., Fox, E. A., and Liu, B., Eds. 2006. *CIKM 2006: Proceedings of the 13th ACM Conference on Information and Knowledge Management*. ACM Press, New York.

Zobel, J. 1998. How reliable are the results of large-scale information retrieval experiments? In *SIGIR'98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, Eds. ACM Press, New York, 307–314.

Received December 2007; revised October 2008; accepted February 2009