Revisiting the Relationship between Document Length and Relevance

David E. Losada
Dep. Electrónica y Computación
Univ. Santiago de Compostela
Spain
dlosada@usc.es

Leif Azzopardi
Dep. Computing Science
Univ. Glasgow
Scotland
leif@dcs.gla.ac.uk

Mark Baillie
Dep. Computer & Inf. Sciences
Univ. Strathclyde
Scotland
mb@cis.strath.ac.uk

ABSTRACT

The scope hypothesis in Information Retrieval (IR) states that a relationship exists between document length and relevance, such that the likelihood of relevance increases with document length. A number of empirical studies have provided statistical evidence supporting the scope hypothesis. However, these studies make the implicit assumption that modern test collections are complete (i.e. all documents are assessed for relevance). As a consequence the observed evidence is misleading. In this paper we perform a deeper analysis of document length and relevance taking into account that test collections are incomplete. We first demonstrate that previous evidence supporting the scope hypothesis was an artefact of the test collection, where there is a bias towards longer documents in the pooling process. We evaluate whether this length bias affects system comparison when using incomplete test collections. The results indicate that test collections are problematic when considering MAP as a measure of effectiveness but are relatively robust when using bpref. The implications of the study indicate that retrieval models should not be tuned to favour longer documents, and that designers of new test collections should take measures against length bias during the pooling process in order to create more reliable and robust test collections.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation, Performance

Keywords

Information Retrieval, Document Length, Relevance, Evaluation, Pooling

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.

1. INTRODUCTION

This paper investigates the assumed relationship between relevance and document length. As highlighted by Robertson and Walker [14], the relationship between document relevance and length can be explained either by: i) *the scope hypothesis*, the likelihood of a document's relevance increases with length due to the increase in material covered; or ii) the *verbosity hypothesis*, where a longer document may cover a similar scope than shorter documents but simply uses more words and phrases.

It is currently believed that the scope hypothesis prevails over the verbosity hypothesis. This belief is supported by a number of empirical studies investigating the relationship between document relevance and length. The probability of a document's relevance (to an information need) is considered to be positively correlated with the length of the document. For instance, Singhal *et al.* illustrated that document relevance increases proportionally with length across a number of early TREC test collections [16]. Similar results have also been reported for later "ad hoc" test collections [10, 9] [1].

Accounting for document length during retrieval was recognized as an important research topic in early indexing models [7]. Recently, it has been reported that retrieval performance can be improved through appropriate term weighting strategies based on document length [6]. Therefore, designing IR models with an a priori preference for longer documents has been viewed as a way to improve retrieval performance [8, 10, 11, 12, 9, 2]. Additionally, a number of studies have proposed adjustments to well known retrieval models in order to account for document length appropriately. For instance, pivoted document length normalization [16] was defined to correct the excessive promotion of shorter documents in the cosine similarity measure within the Vector Space IR model. Similarly, the probabilistic model Okapi BM25 was extended to minimise the bias towards longer documents [15]. The document length correction that is inherent to Statistical Language Models of IR has also been studied in depth [19, 1, 12].

[1] This paper focuses solely on the relationship between relevance and document length in the context of the ad hoc retrieval task. However, for other tasks the relationship between document length and relevance has found to be different. For instance, on the web entry page task in the web collections, there was no correlation found between document length and relevance [11].

Col.	# Docs	# Uniq. Terms	Mean. doc length	Bin size	TREC topics
TREC-2	741856	615087	415	14838	101-150
TREC-3	741856	615087	415	14838	151-200
TREC-4	567493	572107	483	11350	201-250
TREC-5	524929	686358	533	10499	251-300
TREC-6	556077	722085	514	11122	301-350
TREC-7	528155	629880	481	10564	351-400
TREC-8	528155	629880	481	10564	401-450

Table 1: Basic statistics of the TREC adhoc collections, including sizes of the bins for the document length study.

A common theme throughout these studies has been the implicit assumption made that the underlying test collections are complete (i.e. the assumption that all documents in the collection have been assessed for relevance for all topics). However, modern collections are incomplete due to factors related to collection size, cost and effort [6]. In this paper we re-investigate the correlation between relevance and document length in the context of incompleteness across seven TREC ad hoc test collections. We illustrate that the positive correlation between document length and relevance is a consequence of assuming that test collections are complete. When incompleteness is taken into consideration a very different trend is observed: the probability of relevance does not linearly increase with document length. Therefore, the previous trends observed cannot simply be assumed to be a causal link between relevance and document length but potentially an artefact of test collection creation. This finding provides strong evidence refuting the scope hypothesis and the naive heuristic of favouring longer documents for retrieval performance improvement.

The remainder of this paper is as follows. In section 2, we first replicate and extend the original *collection wide* analysis of document length and relevance performed in previous papers [16, 11], to include all available TREC ad hoc collections available. We then go beyond existing analysis by taking into account test collection incompleteness by performing a *pool wide* analysis. This analysis shows distinctive trends providing evidence to refute the hypothesis: a strong correlation exists between relevance and document length. In section 3, we perform an extensive investigation into whether the length bias within the pools affects the comparison between systems in terms of both ranking and performance. Finally, in section 4 we summarize the results of this study and discuss the implications of this work on the comparison, design and evaluation of IR systems.

2. DOCUMENT LENGTH AND RELEVANCE

It has been commonly regarded that longer documents have an increased likelihood of relevance. This is a long held belief stemming from studies such as the one reported in [16], where it was shown that relevant documents tend to be longer than documents within the entire collection. Similarly, in [10, 11], the authors show that the probability of relevance is correlated with document length for a number of adhoc and web TREC collections. In order to study these issues in depth and over time, we consider the seven adhoc collections from the well-established TREC benchmarks (from TREC-2 to TREC-8). We now outline this analysis.

2.1 Experimental Setup

For each test collection, we used the Lemur toolkit[2] to index the collections. Documents were preprocessed using Porter's stemmer [13] but no stop wording was applied. As a result, all lengths reported are based on the count of all the tokens occurring within the documents. To compute the different length patterns (i.e. the distribution of relevance given length), the methodology designed in [16] was adapted as follows. For each test collection, we ordered the documents in the collection by the length and then divided them into equal size bins. Next, we computed the probability of relevance associated to each bin. We set the bin size to be 2% of the collection size (i.e. 50 bins per collection). The statistics about each test collections, the TREC topics and bin sizes are reported in Table 1.

2.2 Collection Wide Analysis

The computation of the relevance pattern in the collection is performed using the corresponding set of relevance judgments available for each TREC adhoc test collection. The number of (query, relevant document) pairs are counted and $p(d \in bin_i | d \; is \; rel)$ is computed as the ratio between the number of pairs that have their document from bin_i, and the total number of (query, relevant document) pairs. This relevance pattern is shown as a solid line (labeled as *Rel.*) in fig. 1. These curves show the trends reported also in [16, 10]: *the probability of relevance grows with the length of the documents*. The general tendency is that the bins with longer documents have a higher probability of relevance.

2.3 Pool Wide Analysis

One might be tempted to infer that relevance evolves as shown in the solid lines in fig. 1. However, caution must be taken when considering these patterns, as these test collections are created through a process known as system pooling. System pooling is used to address the intractability of the completeness assumption [6]. Pooling is a focused sampling of the document collection that attempts to discover all potentially relevant documents with respect to a search topic (e.g. approximate the actual number of relevant documents for a given topic). To do so, a number of (diverse) retrieval strategies are combined to probe the document collection for relevant documents. Each system will rank the collection for a given topic, then the top λ documents from the subsequent ranked lists are collated, removing duplicates, to form a pool of unique documents. All documents in this pool are then judged for relevance by human assessors.

As a consequence, the question we need to ask is: *is the set of documents assessed representative of the collection in terms of length?* For example, if the assessed documents are longer in length relative to the collection then there is an increase in uncertainty concerning the relevance of shorter documents. In this case, the probability of relevance associated to shorter documents may be inaccurate because a smaller proportion of shorter documents will have been assessed for relevance.

To determine whether this is problematic, we counted the number of assessed documents per bin and computed the probability $p(d \in bin_i | d \; is \; judged)$ as the ratio between the number of assessed documents from bin_i, and the total number of assessed documents. This assessment pattern is

[2]http://www.lemurproject.org

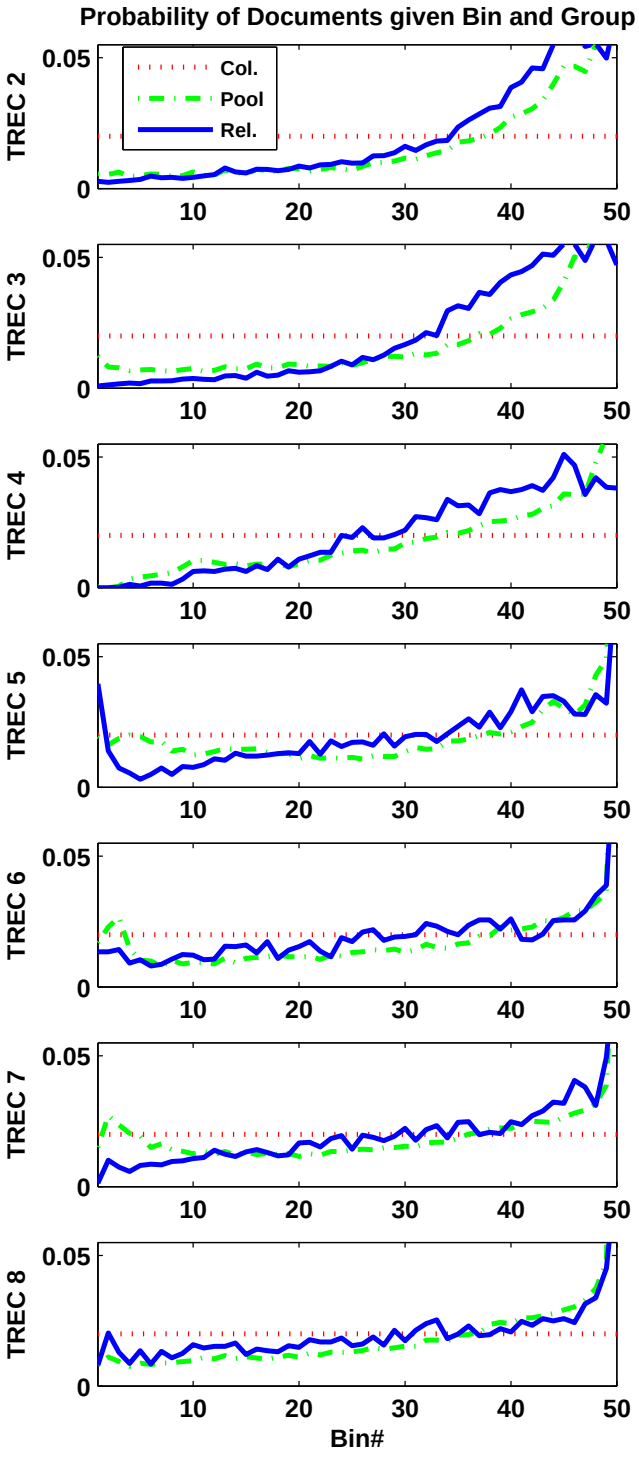

Figure 1: Probability of finding relevant (solid line), assessed (dash-dot line) and collection (dotted line) documents given each bin in the seven TREC adhoc collections. The X axis denotes the bins in increasing order of length. The Y axis denotes the probability of a document being drawn from this bin for a particular group (i.e. the entire collection (Col.), the pool of judged documents (Pool), or the set of relevant documents (Rel.))

presented in fig. 1 as a dash-dotted line (labeled as *Pool*). If the documents were a representative sample from the collection, then we would expect that the number of documents in each bin would be approximately equal. However, the figure indicates that the bins representing longer documents contain more assessed documents in comparison to those bins with shorter documents. Overall, the bins with the longest documents contribute more to the pool of assessed documents. Actually, in terms of the proportion of their contribution there is a very substantial difference between the last few bins and the others. In fact, between 10% and 20% of the assessed documents were found in the last bin. This suggests that the reason for longer relevant documents may be because more longer documents were assessed, as opposed to longer documents being increasingly more likely to be relevant.

The mean and median of document lengths for the collection, pool of assessed documents, set of relevant, and set of (assessed) non-relevant documents are reported in Table 2. To determine the differences between the length distributions, we formally tested whether the distribution of document length within these sets were statistically different using the Mann-Whitney U-test. In all cases, the hypothesis that the documents come from the same distribution was rejected at a 5% significance level. Further, we found that the relevant documents are significantly longer on average than the entire document collection and that the pool of assessed documents contains documents significantly longer on average than both the set of relevant documents and the entire document collection.

Since the pool of assessed documents is not representative of the collection in terms of document length, as it is heavily biased towards longer documents, the finding that longer documents are more likely to be relevant may be an artefact of the pooling process. In other words, the pool of assessed documents is over-represented by longer documents. As a result, there is a higher likelihood of longer documents being judged as relevant, which in turn over-inflates the estimate of the probability of relevance for longer documents.

2.3.1 *Relevance pattern within the pool*

A more appropriate way to estimate a relevance pattern consists of restricting the analysis to the set of assessed documents for each topic. Since all documents in this set are judged, there is no need to take any assumption about the relevance/non-relevance of non-assessed documents.

The probability of relevance within the set of judged documents can be computed as follows. The probability $p(d\ is\ rel | d \in bin_i\ \&\ d\ is\ judged)$ is estimated as the number of (query, relevant document) pairs whose document belongs to bin_i divided by the number of (query, judged document) pairs whose document belongs to bin_i. This analysis is reliable and robust in comparison to the approach reported above. This is because for each bin the relevance counts are divided by the number of assessments in the bin, and therefore, the final curve is not influenced by the skewed distribution of the number of assessments across bins.

The resulting relevance pattern is shown in fig. 2 (right-hand side). For comparison purposes, the relevance pattern computed assuming that non-judged documents are non-relevant, $p(d\ is\ rel | d \in bin_i)$, is also shown in the figure (left-hand side). There is a substantial difference between the trends shown in these two sets of graphs. The left-hand

	Col.		Pool		Rel.		Non-Rel.	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
TREC-2	415	195	6977	747	1443	611	8241	791
TREC-3	415	195	4524	667	1146	584	4902	685
TREC-4	483	266	3774	662	1665	565	3945	674
TREC-5	533	328	3954	542	2857	556	4001	541
TREC-6	514	316	9222	633	2680	482	9668	646
TREC-7	481	329	2304	505	1200	550	2372	502
TREC-8	481	329	2861	676	1129	476	2961	688

Table 2: The mean and median of document lengths in TREC adhoc collections, pools, sets of relevant, and sets of (assessed) non-relevant documents

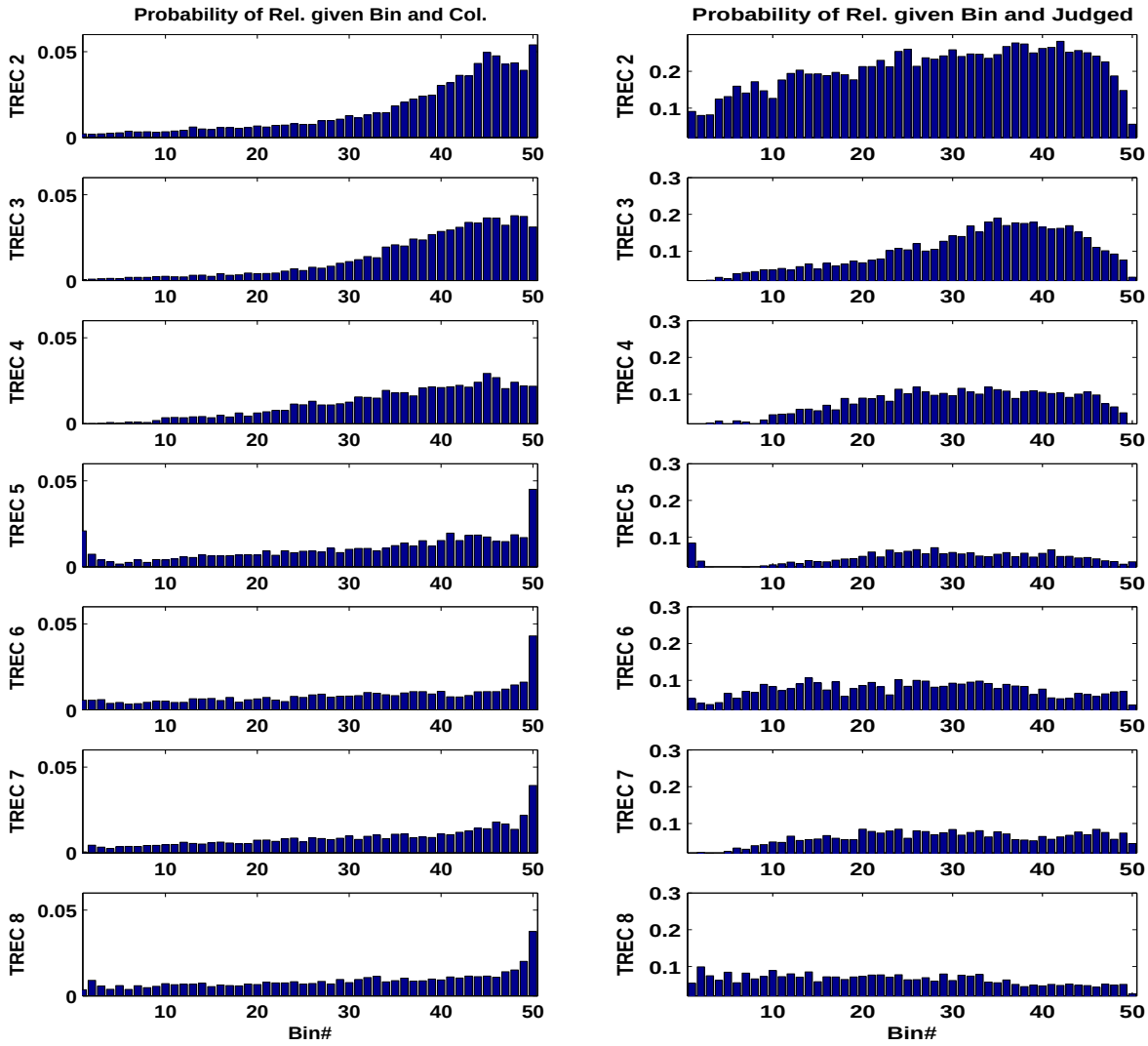

Figure 2: Left: Probability of relevance given bin within the collection. Right: Probability of relevance given bin within the assessed documents. The X axis represents the bin number, ordered by increasing length, and the Y axis represents $p(d\ is\ rel|d \in bin_i)$ and $p(d\ is\ rel|d \in bin_i\ \&\ d\ is\ judged)$, respectively.

graph provides evidence to accept the scope hypothesis, as it clearly shows that probability of relevance increases with length. However, the right-hand graph, which is based on only the assessed documents provides strong evidence to reject the scope hypothesis.

The distribution of relevance within the assessed documents (right-hand graph) tends to be more uniform than the distribution of relevance in the $p(d\ is\ rel|d \in bin_i)$ pattern (left-hand graph). In the right-hand graph, the bins with highest probability of relevance are those containing documents of average length (i.e. bins 20-40) and the probability of relevance falls dramatically for the longest documents (last bins). For the early bins (1-20) the probability of relevance is usually somewhat lower, presumably because a document needs to be large enough to provide sufficient details in order to be relevant. Also, it was observed that the probability of relevance tends to be flatter in the latter TREC collections.

This analysis provides important new insights into the issue of document length and relevance. In particular, this study indicates that previous empirical evidence supporting the scope hypothesis is over-exaggerated. In the next section, we study whether these length biases affect the evaluation of systems that retrieve many short documents.

3. TEST COLLECTION ANALYSIS

In this section we evaluate whether the document length biased pools associated with ad hoc test collections affect evaluation. In other words, do systems that retrieve longer documents on average perform better than systems that retrieve shorter documents? To answer this question, we conduct a thorough analysis on the reliability and robustness for system comparisons under incompleteness.

3.1 The process of Pooling

As explained in section 2, the creation of a complete test collection is impractical [6]. Therefore, only a small focused sample of documents are judged (i.e. assessed) for relevance. The goal of pooling is to maximise the number of relevant document found within a collection while minimising the amount of effort required to perform the assessments. Thus, a small subset of the collection can be assessed in order to obtain a relatively good estimate on the number of relevant documents in the collection.

A fundamental assumption required for pooling is that each run that contributes to the pool is assumed to be independent [17]. If runs are independent then the more runs participating in the pooling the more diverse the relevant documents found are. However, in the construction of the ad hoc TREC collections, the independence assumption has been somewhat relaxed. Many pooled systems implement similar retrieval strategies and, additionally, runs from the same group are likely to show high overlap. This *system reinforcement* bias and the effect of pooling on systems that did not participate in the pool (*system omission*) have been previously evaluated in the literature [20, 18]. It has been shown that the relative order of the systems is quite stable (i.e. the system rankings constructed in decreasing order of a given performance measure such as Mean Average Precision (MAP) are not significantly affected by reinforcement or omission). However, no one has specifically considered whether the length bias within pools affects the evaluation of systems. For instance, it might be the case that pooling is fair to most of the omitted systems (i.e. their position in the rank would not change significantly when they are included into the pool) but it is unfair to a few omitted systems that favor short documents.

In the remainder of this section, we test to determine whether the pools from the ad hoc test collections enable robust comparisons; such that the relative performance of systems that favor longer or shorter documents is not over or underestimated because of the non-representative sample used to form the pools. In order to examine whether the pools provide a robust evaluation against length, we have created different samples of assessments from the official judgments with varying distribution of document length. This helps to study the influence of document length on the evaluation. The relative changes in system rankings and the magnitude of the change in performance scores are carefully analyzed. In particular, we study the correlation between these variations and the length of the documents retrieved by the systems involved.

3.2 Pool samples

In what follows, $p_{off}(.)$ refers to the probabilities computed from the original set of official assessments whereas $p_{sam}(.)$ refers to the probabilities computed from a given sample of the original assessments.

To evaluate how sensitive the relative performance of the systems is with respect to the length of the assessed documents, we created the following samples from the official relevance assessments:

- A sample of the relevance assessments that follows a distribution of lengths as indicated by $p_{off}(d\ is\ rel|d \in bin_i\ \&\ d\ is\ judged)$. The rationale behind this method is that $p_{off}(d\ is\ rel|d \in bin_i\ \&\ d\ is\ judged)$ is a reliable indication of how relevance evolves against document length and, therefore, making that the sampled assessments follow this distribution (instead of taking the whole set of assessments which, as argued above, contains many long documents) we ensure that the bins with more assessed documents are those where the probability of relevance is higher. This can be thought of as sampling method that promotes the document lengths that have higher probability of relevance (rather than simply promoting higher lengths). This basically means that a number of judgments associated to long documents are removed from the original assessments and the final shape of the relevance curve in the new assessments ($p_{sam}(d \in bin_i|\ d\ is\ rel)$) resembles the right-hand pattern shown in fig. 2. We therefore simulate a situation in which the input to the judgment process is more uniform with respect to document length and, thus, the output of the judgment process is not strongly biased towards high lengths. This sample will be referred to as `towards_prel_in_pool`.

- A sample from the original assessments where the top 25% longest documents are removed. This sample will be referred to as `long_removed`.

- A sample from the original assessments where the top 25% shortest documents are removed. This sample will be referred to as `short_removed`.

- A sample from the original assessments where the top

Collection	# Original Pool	# Sampled Pool tpip	lr,sr	tr
TREC-5	133682	46805	100261	66841
TREC-6	72271	23241	54203	36135
TREC-7	80346	32688	60259	40173
TREC-8	86831	29058	65123	43415

Table 3: Size of the original assessments vs Size of the sampled assessments: `towards_prel_in_pool` (`tpip`), `long_removed` (`lr`), `short_removed` (`sr`) and `tails_removed`(`tr`)

25% shortest and top 25% longest documents are removed. This sample will be referred to as `tails_removed`.

These samples help to study the effect of document length bias on the relative comparison of systems: each sample simulates a pooling scenario related to a specific form of length bias. The influence of this bias was then evaluated over a large number of system comparisons.

3.2.1 *Sample Creation*

The last three samples are straightforward to construct. We ordered the documents in the pool in decreasing order of length and the top 25% documents are removed (`long_removed`), the bottom 25% documents are removed (`short_removed`) or both the top 25% documents and the bottom 25% documents are removed (`tails_removed`). This removal process was done globally rather than on a query-by-query basis[3].

To obtain the `towards_prel_in_pool` sample we first compute the sample size. This is determined by the number of assessments in the original pool, the distribution of assessments across the bins and the distribution $p_{off}(d\ is\ rel|d \in bin_i\ \&\ d\ is\ judged)$ (which is the probability distribution we want the sample to follow). Next, we draw randomly documents from the original assessments, ensuring that each bin obtains the required number of assessments.

The new sample of assessed documents now reflects a pattern similar to the relevance pattern within the set of original assessments. More specifically, when taking these sampled judgments to estimate the probability of relevance within the collection as a whole (i.e. assuming that non-judged documents are non-relevant) one can observe that the shape of the curve reflects now the probability pattern within the assessed documents. This is shown in fig. 3, whose relevance pattern resembles the one shown in fig. 2 (right-hand graph). This means that the new set of judgments is more representative of the relationship between relevance and document length.

The size of the original assessments and the size of all these sampled assessments are reported in Table 3.

3.3 The stability of system rankings

In this section we check whether or not the measured relative performance of TREC participants is reliable. As argued above, one might be tempted to think that systems biased towards short documents could be harmed by the high population of long documents in the TREC pools. For this part of the study, we used the pooling runs submitted to the TREC adhoc task from TREC-5 to TREC-8, which were provided by NIST[4]. Given the official relevance assessments and the four sets of assessments reported in the last section, we computed the system rankings using both bpref and MAP. Unlike MAP, bpref does not assume that non-assessed documents are non-relevant but estimates performance using the judged set of documents only [4].

The association between the different system rankings was quantified applying Kendall's tau. A similar method was taken in [18], where Voorhees evaluated the stability of system rankings with respect to judgments created by different sets of human assessors. Kendall's tau is defined as the minimum number of transpositions required to turn one ranking into the other. This number is normalized such that two identical rankings produce a correlation equal to 1, whilst the correlation between a ranking and its inverse is equal to -1. The correlation results are presented in Table 4. The table reports the correlation between the ranking produced from the official assessments and each ranking

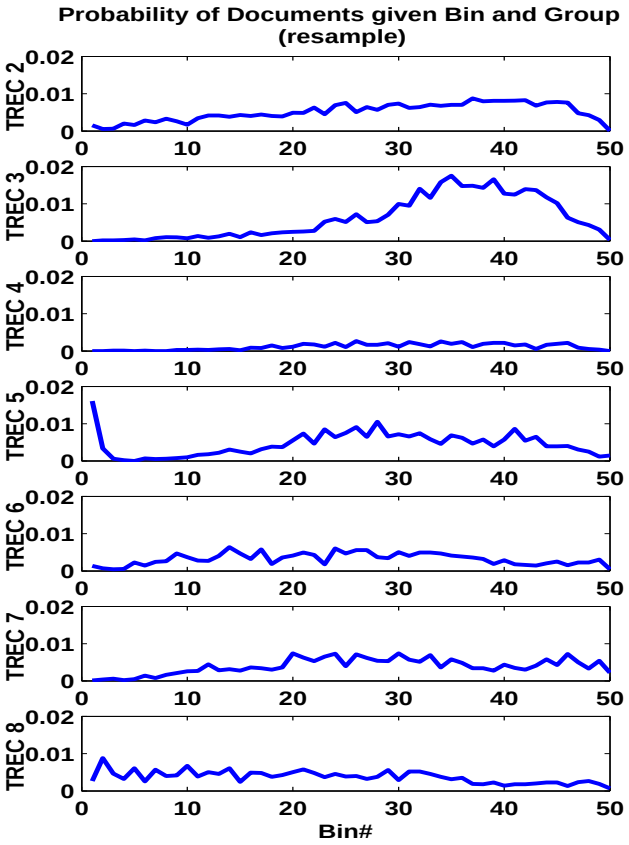

Figure 3: Probability of relevance given bin computed from the `towards_prel_in_pool` sampled assessments. The X axis represents the bin number and the Y axis represents $p(d \in bin_i|d\ is\ rel)$.

[3]We also conducted the same experiments following a query-by-query sampling and the trends and conclusions found are the same as those discussed here.

[4]NIST could not provide the information on the pooled runs for earlier TREC years. Therefore, we focus our pooling analysis on TRECs 5 to 8.

		bpref		MAP	
	off vs	Corr.	p-value	Corr.	p-value
TREC-5	tpip	0.9299	$\approx 10^{-43}$	0.8665	$\approx 10^{-37}$
(101 runs)	sr	0.9374	$\approx 10^{-44}$	0.9026	$\approx 10^{-41}$
	lr	0.9081	$\approx 10^{-41}$	0.8887	$\approx 10^{-39}$
	tr	0.9065	$\approx 10^{-41}$	0.8879	$\approx 10^{-39}$
TREC-6	tpip	0.8763	$\approx 10^{-16}$	0.8396	$\approx 10^{-16}$
(46 runs)	sr	0.9014	$\approx 10^{-16}$	0.8783	$\approx 10^{-16}$
	lr	0.9169	$\approx 10^{-16}$	0.8860	$\approx 10^{-16}$
	tr	0.8628	$\approx 10^{-16}$	0.8493	$\approx 10^{-16}$
TREC-7	tpip	0.9214	$\approx 10^{-35}$	0.8411	$\approx 10^{-29}$
(84 runs)	sr	0.9036	$\approx 10^{-34}$	0.8927	$\approx 10^{-33}$
	lr	0.8646	$\approx 10^{-31}$	0.8371	$\approx 10^{-29}$
	tr	0.8600	$\approx 10^{-31}$	0.8663	$\approx 10^{-31}$
TREC-8	tpip	0.9042	$\approx 10^{-29}$	0.6893	$\approx 10^{-17}$
(71 runs)	sr	0.9050	$\approx 10^{-29}$	0.8680	$\approx 10^{-26}$
	lr	0.8221	$\approx 10^{-24}$	0.7642	$\approx 10^{-21}$
	tr	0.8592	$\approx 10^{-26}$	0.7755	$\approx 10^{-21}$

Table 4: Kendall's tau correlation between the official system rankings and the system rankings obtained with the sampled assessments: `towards_prel_in_pool` (`tpip`), `long_removed` (`lr`), `short_removed` (`sr`) and `tails_removed`(`tr`).

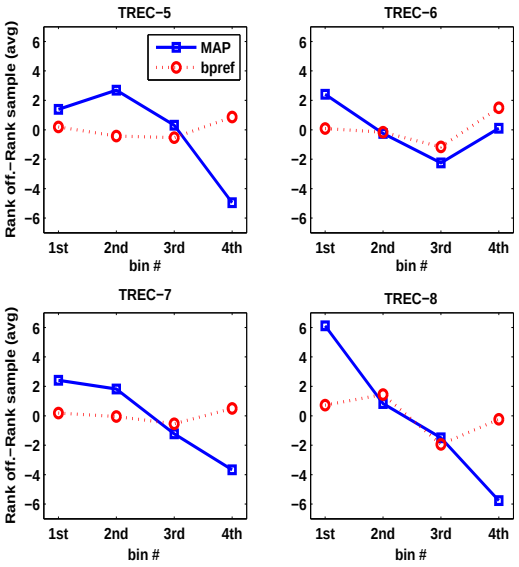

Figure 4: Difference between the rank obtained by a system with the official assessments and the rank obtained with the `towards_prel_in_pool` sampled assessments. The figures refer to the mean difference computed across the runs in the bin.

produced from every sample of assessments and every performance measure. Additionally, the table informs about the p-values obtained for testing the hypothesis of no correlation. The results are very conclusive. For all collections and pairwise comparisons, there is a very high correlation between the official system rankings and the rankings produced from different length-biased samples. This provides evidence to show that the official rankings are relatively insensitive to the distribution of document lengths in the pools of assessed documents. This is good news to current IR evaluation standards as it shows that, although there is a bias in favour of long documents in these pools, this bias does not significantly affect the general conclusions drawn in TREC reports.

Observe also that the correlations computed with MAP tend to be lower than the correlations computed with bpref. This demonstrates that bpref is less sensitive to the relevance judgments utilized than MAP: the resemblance between the bpref system rankings computed from the samples and the official bpref rankings is higher than the resemblance between the MAP system rankings computed from the samples and the official MAP rankings. This is evidence to support bpref as a robust measure to deal with incomplete judgments.

These correlation values demonstrate clearly that there are only minor changes in the systems rankings. Still, it could be the case that the few systems that obtain a substantially different rank are somehow affected by its retrieval trends against document length. To further check this issue, we ordered the runs in increasing order of the average length of the documents retrieved in the top 100. Next, the runs were divided into four bins: the first bin contains the runs that retrieve shorter documents while the last bin contains the runs that retrieve longer documents. For each bin, we computed the average difference between the system rank obtained from the official judgments and the system rank obtained with the `towards_prel_in_pool` sample. If this difference is positive then the system was promoted by the sample. In contrast, a negative value means that the system was demoted by the sample. This permits to analyze whether or not there is any promotion/demotion tendency against the length retrieval behaviour of the runs. The results are reported in fig. 4. Two main observations can be made here. First, the system rank computed using bpref is quite stable across bins, meaning that there is not any tendency to promote or demote runs that retrieve either shorter or longer documents. Second, with MAP, the runs that retrieve shorter documents (bin 1) show a clear tendency to be promoted by the sample, while the runs retrieving longer documents (e.g. bin 4) tend to be demoted by the sample. This means that the official judgments tend to underrate (overrate) the runs that retrieve shorter (longer) documents when used to rank systems with MAP. Although the correlations reported above show that the rankings with the official judgments and the rankings with the sampled judgments are quite similar, we found here that the variations are not distributed uniformly across runs but there is a tendency to harm runs that retrieve shorter when the official judgments are used. This is strong evidence to reject MAP as a performance measure to rank systems because MAP, together with the biased pools, establish a preference for particular types of systems. In contrast, bpref handles well the incompleteness of the judgments and treats fairly the runs that retrieve shorter documents.

3.4 System omission

Having analyzed pooling for the systems that contributed to the pools, it is also interesting to analyze the effect of length for runs that did not have an opportunity to contribute to the pool. A given test strategy might be reliable for relative comparison of the systems that participated in the pooling process but, in contrast, the collection might be not reusable because it does not handle fairly the *non-pool*

bpref improvements

	official vs LOU				official vs LOUG			
	mean	max	min	std dev	mean	max	min	std dev
TREC-5	+1.45%	+50%	+0.0%	6.4	+2.34%	+66%	−0.05%	8.79
TREC-6	+2.32%	+46.67%	+0.0%	6.91	+2.44%	+46.67%	+0.0%	6.88
TREC-7	+0.25%	+3.19%	−0.13%	0.47	+0.50%	+6.47%	−0.13%	1.07
TREC-8	+0.61%	+7.03%	+0.0%	1.34	+0.75%	+7.03%	+0.0%	1.56

MAP improvements

	official vs LOU				official vs LOUG			
	mean	max	min	std dev	mean	max	min	std dev
TREC-5	+0.35%	+3.02%	+0.0%	0.57	+0.91%	+5.21%	+0.0%	1.31
TREC-6	+2.13%	+33.3%	+0.0%	5.07	+2.32%	+33.3%	+0.0%	5.07
TREC-7	+0.34%	+5.36%	+0.0%	0.76	+0.54%	+5.36%	+0.0%	0.95
TREC-8	+0.59%	+5.61%	+0.0%	1.21	+0.82%	+10.49%	+0.0%	1.96

Table 5: Improvement in performance computed as the percentage difference between bpref/MAP on the official pool and bpref/MAP on the LOU/LOUG modified pool. The table includes the mean, the maximum and the minimum improvement obtained over all pool runs and the standard deviation.

runs.

To investigate this issue, we adopt the methodology proposed by Zobel [20], which was later referred to as *Leave Out Uniques* (LOU) [3]. Each run that contributed to the pool is evaluated first against the official assessments and, next, the same run is evaluated using the pool with the documents contributed only by the run removed. The latter evaluation simulates the situation where the run did not participate in the pools. The difference in the evaluation ratios between both cases (averaged over all runs) is a measure of the degree to which contributing to the pool improves effectiveness. This measurement is a natural way to check the reusability of a given collection.

Since runs from the same group tend to show a high overlap in the retrieved documents, it is also interesting to test an alternative to the LOU test that consists of removing all documents from the judgment set that were contributed solely by runs from the same organization. This method was proposed in [3] as a more stringent variant of the LOU test. In the following, LOU refers to the original test as proposed by Zobel whereas LOUG (*Leave Out Uniques Group*) stands for the group-oriented variant.

Table 5 presents the results obtained with bpref and MAP for both LOU and LOUG test. It is interesting to observe that there are not major differences here between bpref and MAP. Both measures show small improvements in performance, similar to those reported in [20]. This supports the belief that TREC collections are reusable because, on average, participating in the pool does not produce a major benefit in terms of performance.

It is interesting to observe that TREC-7 and TREC-8 appear more reliable because the average improvements are smaller in these collections. This can be explained as follows. First, the mean length of the documents in the pools is lower in TREC-7 and TREC-8 than in TREC-5 and TREC-6 (refer to Table 2). Second, if we consider the ratios between the average lengths[5] of the pool and collection for each of the successive TREC years, we can consider the trend over time. Figure 5 illustrates that the ratio tends to decrease with each new TREC collection. This observation would appear to indicate that over time those IR systems contributing to the pooling process compensate more appropriately for document length, such that pools from the latter TREC collections are less biased in terms of length and more representative of the collection.

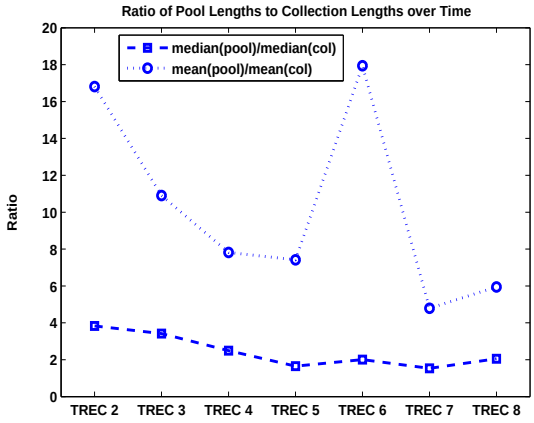

Figure 5: Evolution across the years of the ratios of mean and median document length in collection and pool.

Returning to the main point of our analysis, we were interested in analyzing carefully the improvements in performance against these document length retrieval trends. To do so, we grouped the system runs into bins using the same strategy explained in the previous section. Figure 6 plots the relative improvement in performance (averaged across the runs in the bin) obtained from the participation in the pool against the bin number. The graph includes a plot for each combination of performance measure and type of leave-out test. In this way, we can check whether or not there exists any length effect. If a given set of non-pool runs promoting short documents was unfairly penalized by the of-

[5] $avg(col)$ ($median(col)$) is the average (median) document length in the whole collection, while $avg(pool)$ ($median(pool)$) is the average (median) document length of the assessed documents.

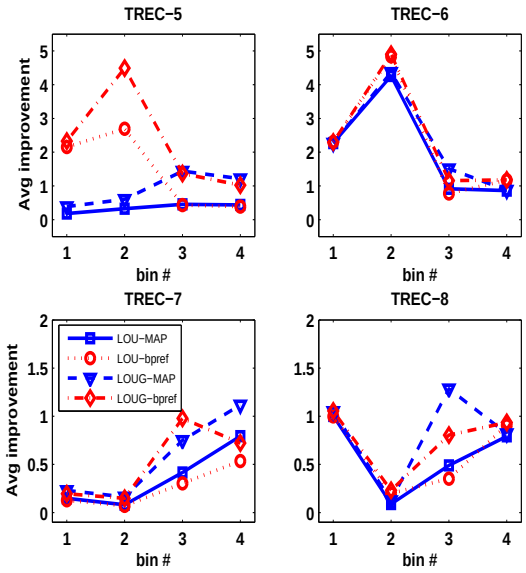

Figure 6: Improvement in performance computed as the percentage difference between MAP/bpref on the official pool and MAP/bpref on the LOU/LOUG modified pool. The figures refer to the mean improvement computed across the runs in the bin.

ficial judgments then it would get substantial improvements when included in the pools. Conversely, non-pool runs promoting long documents would obtain little benefit because current pools are mostly populated by long documents. In the figure, we can observe that the improvements are tiny and relatively uniform across bins in TREC-7 and TREC-8. In contrast, in TREC-5 and TREC-6, the runs retrieving shorter documents (bins 1 and 2) tend to present higher improvements. This indicates that, in these collections, there is a disproportionate tendency to underrate non-pool runs that retrieve shorter documents. In conjunction with the trend analysis above, this provides supporting evidence to suggest that the more representative the pool is in terms of length the more reusable the collection will be.

4. DISCUSSION AND CONCLUSION

This paper provides statistical evidence to reject a commonly held assumption made in the IR literature [16, 10, 9, 8, 2], the scope hypothesis: the hypothesis that the probability of relevance increases with document length in ad hoc retrieval. The link between length and relevance has been based on a series of empirical studies. These studies make the implicit assumption that test collections are complete, overlooking the methods used to construct the test collections such as system pooling. As a result, the presence of longer documents returned from pooling (relative to the average document length in the collection), strongly influences the shape of the relevance pattern.

In the context of test collections, this investigation illustrated that for all TREC ad hoc collections, the pool of assessed documents have a larger frequency of longer documents in comparison to the test collection. However, the probability of relevance associated with the longest set of documents is lower than the probability of relevance associated to the smallest set. This is an indication that systems forming the assessment pool bias towards longer documents. The pool is therefore a biased sample of the collection. More importantly, the assessment pool of documents is not representative when considering the sample of relevance and judged documents. Additionally, the probability of relevance given document length is more uniform in the later collections, where the bias towards longer lengths is less extreme. This is an important result indicating that relevance is not strongly associated with document length.

Therefore, it was crucial to assess whether these length-biased pools were problematic for comparing IR algorithms. Here, we showed that the rankings of the systems participating in TREC were not significantly affected by such bias. However, our study did indicate that when using MAP to estimate system performance, the performance of those systems retrieving longer documents tended to be overestimated, while the performance of those system retrieving shorter documents was underestimated. In contrast, the system rankings computed using bpref are quite insensitive to this document length bias. This highlights an important distinction between bpref and MAP. In comparison to measures such as MAP, bpref does not make the assumption that those documents not assessed are not relevant. Instead, bpref estimates performance using only the judged set of documents thereby minimising potential bias, and providing a more robust IR measure.

A further general finding from the study indicated that the TREC-7 and TREC-8 collections appear to be more reusable than TREC-5 and TREC-6. This is because the earlier collections tend to underestimate significantly the performance of those systems that were not included within the pooling process which retrieve shorter documents on average. As stated previously, to minimise this potential system omission bias, metrics that account for incompleteness, such as bpref, should be considered over MAP.

In the context of retrieval algorithms, this study indicated that the probability of relevance does not necessarily increase with document length. Therefore, the assumption that longer documents are more likely to be relevant, implicit in many retrieval models, should be reconsidered. For example, a model that sets a document prior that grows increasingly with document length will appear to obtain better performance. However, this is because the model is overfitted towards the set of length-biased relevance assessments. As a result, the retrieval model may have limitations when generalising to other topics, collections, or even to updates in the collection.

This paper opens new research lines into IR evaluation techniques. We argue that this bias in the pooling process, towards longer documents, should be corrected. Instead of simply taking the union of the top λ retrieved documents from a series of contributing systems, the pools could be constructed taking into account the shape of the relevance pattern against document length. In other words, form a representative sample taking into account the actual relevance pattern against document length. This line of research is similar to the one followed in [20, 5], where the authors designed new pooling methods that find more relevant documents in fewer total documents judged. In [20], Zobel applied variable-depth pooling to judge more documents for topics that are predicted to have many relevant documents.

In [5], the authors suggest to insert more documents into the pools from the runs that returned more relevant documents recently. However, none of these studies consider any document length policy. Our future work will consider these issues.

Finally, another interesting line of research consists of conducting a real world study into length and relevance and analyzing what kind of relationship exits outwith the laboratory setting e.g. within the context of an information seeking study.

5. ACKNOWLEDGMENTS

This work was partially supported by projects TIN2005-08521-C02-01 (Ministerio de Educación y Ciencia, Spain) and PGIDT07SIN005206PR (Xunta de Galicia, Spain). We thank the TREC project managers for providing us the retrieval runs needed to conduct our study.

6. REFERENCES

[1] L. Azzopardi and D. Losada. Fairly retrieving documents of all lengths: A study of document length normalization using the language modeling approach. In *Proc. 1st International Conference on the Theory of Information Retrieval, ICTIR'07*, pages 65–75, Budapest, October 2007.

[2] R. Blanco and A. Barreiro. Probabilistic document length priors for language models. In *Proc. ECIR-08, the 30th European Conference on Information Retrieval Research*, pages 394–405, Glasgow, United Kingdom, March 2008.

[3] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees. Bias and the limits of pooling for large collections. *Information Retrieval*, 10:491–508, 2007.

[4] C. Buckley and E. Voorhees. Retrieval evaluation with incomplete information. In *Proc. SIGIR-04, the 27th ACM Conference on Research and Development in Information Retrieval*, pages 25–32, Sheffield, UK, July 2004.

[5] G. Cormack, C. Palmer, and C. Clarke. Efficient construction of large test collections. In *Proc. of SIGIR-98, the 21st ACM International Conference on Research and Development in Information Retrieval*, pages 282–289. ACM press, August 1998.

[6] D. Harman. *TREC:Experiment and Evaluation in Information Retrieval*, chapter The TREC AdHoc Experiments, pages 79–97. The MIT press, 2005.

[7] S. Harter. A probabilistic approach to automatic keyword indexing. *Journal of the American Society for Information Science*, 26:197–206, 1975.

[8] D. Hiemstra. A probabilistic justification for using tf x idf term weighting in information retrieval. *Int. Journal of Digital Libraries*, 3:131–139, 2000.

[9] J. Kamps. Web-centric language models. In *Proc. ACM Conference on Information and Knowledge Management (CIKM)*, 2005.

[10] W. Kraaij and T. Westerveld. Tno/ut at trec-9: how different are web documents. In *Proc. TREC-9, the 9th Text Retrieval Conference*, Gaithersburg, United States, November 2000.

[11] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *Proc. 25th ACM Conference on Research and Development in Information Retrieval, SIGIR'02*, pages 27–34, Tampere, Finland, 2002.

[12] D. Losada and L. Azzopardi. An analysis on document length retrieval trends in language modeling smoothing. *Information Retrieval*, 11:109–138, 2008.

[13] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[14] S. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proc. SIGIR-94, the 17th ACM Conference on Research and Development in Information Retrieval*, pages 232–241, Dublin, Ireland, July 1994.

[15] S. Robertson, S. Walker, S. Jones, M. Hancock Beaulieu, and M. Gatford. Okapi at TREC-3. In D.Harman, editor, *Proc. of the TREC-3, the 3rd Text Retrieval Conference*, pages 109–127. NIST, 1995.

[16] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proc. of the 19th ACM SIGIR conference on Research and Development in Information Retrieval*, pages 21–29, 1996.

[17] K. Sparck Jones and C. J. van Rijsbergen. Report on the need for and provision of and ideal information retrieval test collection. Technical report, British Library Research and Development Report, 1975.

[18] E. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36:697–716, 2000.

[19] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 334–342, 2001.

[20] J. Zobel. 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*, pages 307–314. ACM Press, 1998.