Differences in Effectiveness Across Sub-collections

Mark Sanderson[♯] Andrew Turpin[Ψ] Ying Zhang[♯] Falk Scholer[♯]

[♯]School of Computer Science & IT
RMIT University
Melbourne, Australia
<firstname.lastname>@rmit.edu.au

[Ψ] Department of Computing and Information Systems
University of Melbourne
Melbourne, Australia
aturpin@unimelb.edu.au

ABSTRACT

The relative performance of retrieval systems when evaluated on one part of a test collection may bear little or no similarity to the relative performance measured on a different part of the collection. In this paper we report the results of a detailed study of the impact that different *sub-collections* have on retrieval effectiveness, analyzing the effect over many collections, and with different approaches to sub-dividing the collections. The effect is shown to be substantial, impacting on comparisons between retrieval runs that are statistically significant. Some possible causes for the effect are investigated, and the implications of this work are examined for test collection design and for the strength of conclusions one can draw from experimental results.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (effectiveness)*

General Terms

Experimentation, Measurement, Performance

Keywords

Search engines, information retrieval evaluation, sub-collections

1. INTRODUCTION

A convention of search engine evaluation is that testing is only rigorous when it is conducted on multiple collections. The ranking function in system A can only be said to be better than the function in system B if A is measured to be more effective than B across all or nearly all tested collections. The early TREC *ad hoc* collections have this convention built into them by virtue of their being composed of multiple *sub-collections*, each different from the others. An aim of this convention is to create robust generic ranking algorithms. However, although testing on multiple collections is established, it has not been researched in much detail. This paper represents an initial investigation into these issues, considering the following research questions:

- Do generic ranking functions search consistently across the distinct sub-collections of a test collection?
- Do different ways of dividing a collection lead to similar inconsistencies?
- Are any inconsistencies found in effectiveness meaningful?
- What are the causes of such inconsistency?

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
Copyright 2012 ACM 978-1-4503-1156-4/12/10...$15.00.

2. RELATED WORK

Ever since test collections were first described [1] there has been research conducted on the reliability of results measured from them. Areas of focus have included the extent to which assessors could be trusted to generate reliable objective query relevance judgments (*qrels*) [2], and the number of topics needed in a test collection in order for the ranking to be reliable [3].

One aspect of test collections that has not been examined in as great a detail, is the makeup of the collection. Tague-Sutcliffe [4], in her "*pragmatic guidelines*" for IR experimentation, discussed the selection of document collections, but did not cover the importance of testing on multiple collections. More recently, Soboroff [5] stated "*If significant and noticeable differences in effectiveness are observed … across multiple test collections, we can conclude that one search algorithm is better than another*". Although one can find research that uses this approach as a means of validating results, research that studies how one selects such multiple collections appears to be missing.

The early TREC *ad hoc* collections could be viewed as a set of test collections (called here *sub-collections*), each sourced from different providers [6]. In the first few years of experimentation, large differences in average document length found across the sub-collections meant that ranking functions behaved erratically. However, with the introduction of the length normalization of BM25 [7] and similar techniques in the SMART system [8], research groups started to overcome these challenges by around the TREC-4. Consequently, variation in effectiveness across the TREC source sub-collections was not as much of a focus for IR researchers as it was earlier.

Scholer et al. [9] reported the result of an experiment that split the TREC *ad hoc* test collections into two sub-collections, and compared the effectiveness ranking of retrieval systems on each using Kendall's Tau (τ). Across eight TREC *ad hoc* collections, the correlation between the two ranks of runs ranged from 0.4-0.7. It was thought the low correlation was due to changes in the way that test collection assessors judged documents. However, it was realized that the low τ might be due to inconsistencies in the way the documents from different sources were ranked.

3. METHODOLOGY

For this preliminary study on the consistency of search across sub-collections, TREC run data was used to provide a wide range of collections and retrieval systems to test on. We used the runs submitted to the *ad hoc* tracks in TRECs 2-8 and the terabyte track in TRECs 2004-2006 (Gov2 [10]). Each set of runs represents a variety of different retrieval approaches. We use only automatic runs in our analysis and follow common practice in this style of experiment by discarding the bottom 25% of submitted runs [11], to remove 'buggy' IR systems from the comparisons.

3.1 Sub-collection splitting

Sub-collections were formed using the following criteria.

Primary source: Much of the earlier TREC data (TREC 2-8) is from edited sources (newspapers, government documents, etc.). The source can be easily identified from the TREC-assigned document identifier. This allows us to investigate whether documents from different sources are retrieved differently.

Top-level domains: The Gov2 data set contains 25,205,179 documents from 17,186 distinct hosts. Within this collection '.gov' and '.us' are the dominant TLDs. As with the primary source in TREC *ad hoc*, we decided to test if documents from different domains resulted in different retrieval characteristics.

MIME content types: Another way of splitting the Gov2 collection is by document type. The collection holds 23,111,957 documents of type *text/html* (91.7%), 2,030,339 of type *application/pdf* (8.1%), and a few other document types.

3.2 Splitting a collection

The focus of this paper is on retrieval from different types of sub-collections. The TREC run data contains at most the top 1,000 documents retrieved over whole collections, not the sub-collections we were interested in. We hypothesized that it would be possible to simulate search over sub-collections by partitioning the documents in the runs and qrels of a test collection according to the sub-collection criteria listed above. This was the approach used by Scholer et al. [9]. However, the authors of that paper did not report any testing of the validity of the approach.

Given a document collection, a set of runs, a set of *qrels*, and chosen criteria to split a collection into sub-collections, the ranking of runs measured on each sub-collection was conducted as follows:

1. The document collection was partitioned into sub-collections based on the chosen method.
2. For each sub-collection, new *qrels* were formed, containing only documents from that sub-collection.
3. For each sub-collection, a new set of runs was formed, with each run only containing the documents retrieved from that sub-collection. Documents from outside the sub-collection that occurred in the original run were removed, and eligible items were promoted up the ranked list to fill empty slots.
4. For each newly formed sub-collection run, the Mean Average Precision (MAP) of the run was computed using its respective sub-collection *qrels*.
5. For each sub-collection, the newly formed runs were ranked by their MAP score.
6. For every possible pairing of sub-collections, the Kendall's τ between the run rankings was computed.

If there was no effect from sub-collection splitting, the ranking of systems would be the same on each sub-collection: that is, the top system would always be the top system, regardless of the sub-collection being searched, and so on for all other systems. However, it was not expected that the measure of τ between the ranks of runs would show perfect agreement, since the *qrels* were different for the two sub-collections, introducing a level of noise in the correlation between the rankings. In order to understand the impact of this noise, we conducted a further comparison by splitting the document collection into *randomly* formed sub-collections multiple times. The τ measured between a rank of runs measured between randomly formed sub-collection was collated to produce a distribution of τ values, enabling a randomization test [12] to be carried out. Random splitting for each sub-collection was performed as follows:

System	Ranking function
Indri	Dirichlet, Jelinek-mercer, Okapi
Terrier	DER_BM25, DLH13, In_expB2
Zettair	Dirichlet, Okapi, Pivoted-cosine

Table 1: The IR systems and ranking functions used.

1. For each sub-collection formed in step 1 above, measure the total number of documents in the sub-collection.
2. Randomly select a set of documents from the entire collection up to the measured size of that sub-collection.
3. For each randomly generated sub-collection, repeat steps 2 through 6 in the process described above.

For every possible pairing of sub-collections, the random splitting process was repeated 1,000 times, giving a distribution of τ values based on the pair's random counterparts. The value of τ computed for the actual pair of sub-collections was compared against this distribution, enabling the calculation of a *p*-value.

3.3 Validating the splitting methodology

To test the sub-collection simulation process, we conducted a set of experiments on the TREC-8 collection using three open-source IR research platforms Indri, Zettair, and Terrier. Each was set up to run three different ranking options (see Table 1), effectively providing us with nine different IR systems.

The systems in their different configurations were set to index and retrieve over the full TREC-8 test collection. Using steps 1-6 described above, simulated sub-collection runs were also formed and the MAP of the nine IR systems was measured on those sub-collections. These MAP scores were labeled *simulation*. The IR systems were then set to index and retrieve over each of the sub-collections in isolation. From these, a set of *true* MAP scores were produced. The MAP scores of the nine true systems and the simulations match very closely, with a Pearson correlation of 0.995, demonstrating the validity of splitting existing run data to simulate sub-collections.

4. TREC NEWSWIRE COLLECTIONS

The runs from the newswire *ad hoc* collections of TRECs 2-8 were split into their respective sub-collections, based on primary source, giving from 6 to 15 sub-collections for each year, and 68 possible pairings of sub-collections across the 7 years. To illustrate the data gathered, details of the comparisons made between the sub-collections of TREC 8 are shown in Table 3. Here, the Kendall's τ for the ranking of systems on each sub-collection pair is shown, along with the range of τ values for the 1,000 comparisons of random sub-collections of the same size as the pair, together with the *p*-value from the randomization test. As can be seen, for all but one pair of sub-collections, the value of τ was lower than that for any of the random pairings.

Across all years of TREC, the average Kendall's τ between the system ranks was 0.52, substantially lower than the thresholds usually considered as acceptable when considering the stability of retrieval results measured on test collections.

As detailed in Section 3.2, the 1,000 random pairings of sub-collections conducted for each of the 68 comparisons formed a distribution from which a randomization test could be conducted. Details of the test results are summarized in Table 2. Across the 68 sub-collection pairings, 55 of the measured τ values were found to be statistically significant at the 0.05 level (81%), 48 at the 0.01 level (71%) and 36 at the 0.001 level (53%). If the difference between the sub-collections had simply been due to

	Comparisons significant at		
TREC	*p*≤0.05	*p*≤0.01	*p*≤0.001
8	100%	100%	83%
7	100%	83%	83%
6	70%	40%	20%
5	73%	67%	33%
4	60%	47%	20%
3	100%	100%	100%
2	100%	100%	100%

Table 2: Testing significance using randomization test.

random chance, the expected percentages for the three levels would be respectively 5%, 1% and 0.1%. One can conclude that if a set of retrieval systems were ranked using a source-based sub-collection from TREC, one cannot be confident that the ordering of systems will be the same or similar if tested on a different sub-collection.

We note that τ is affected by all entries in a list, and as such might be influenced by the somewhat arbitrary ranking of statistically similar systems within runs. For example, if the top five systems in a run are only different due to chance (random sampling of topics), and a permutation of the run has the same systems in the top five positions but in a different order, then the τ value between this run and the permutation will be less than one, although there is no practical difference between the two orderings. However, past work on test collection measurement consistency has used similar data sets and methodologies. For example, Voorhees' study of *qrel* assessor error used the same TREC *ad hoc* collections and runs. One might therefore expect the same low τ values due to highly similar run re-orderings; however, Voorhees reported much higher τ values than those observed here [2]: across the experimental results described τ ranged between 0.84 and 1.

Another possible cause of the low τ is the possibility that particular sub-collections will not have any relevant documents available for some topics. We examine this issue in Section 6.3 and show that it is not a factor.

As already described, the large differences in document length found across the sub-collections of TREC *ad hoc* were a challenge for many early ranking functions, but were thought to be solved with the introduction of new ranking functions about mid-way through the running of TREC *ad hoc*. However, if the low τ values were due to the early problems of dealing with the length of documents, one might expect τ to increase in later years. Therefore, the number of rank comparisons that were statistically significant was examined across the years of TREC *ad hoc* to determine if there was any discernible trend. However, as can be seen in Table 2, no trend over time is apparent.

Within TREC, some of the sub-collections were from general news sources, namely the LA Times, Financial Times, San Jose Mercury News, Wall Street Journal, and Associated Press. Since these sub-collections may therefore be considered as a natural grouping, we examined their behavior separately. There were eleven sub-collection pairings between these sources; the average τ between the ranks of runs on this news data was 0.77, notably higher than the τ of 0.52 measured across all 68 pairings. Only five of the τ values across the eleven pairings were significant at the 0.05 level (i.e. 45%), a substantially smaller proportion than that measured across the full set of 68, though still more than would be expected by chance (i.e. 5%).

For the *ad hoc* TREC collections, this was the only discernible indication that ranking systems on these sub-collections might be stable; for all the other sub-collections there was a correlation between the rankings of systems that was statistically significantly lower than would be expected by chance. That is, the system rankings measured on one sub-collection appear to be unique to that sub-collection.

TREC-8		Random range		
Pair	τ	from	to	*p*-value
FBIS-FR	0.46	0.51	0.85	< 0.001
FBIS-FT	0.58	0.65	0.87	< 0.001
FBIS-LA	0.55	0.61	0.85	< 0.001
FR-FT	0.26	0.55	0.85	< 0.001
FR-LA	0.34	0.53	0.84	< 0.001
FT-LA	0.67	0.65	0.87	0.003

Table 3: Comparisons for the TREC-8 sub-collections.

These results confirm that the principle of testing on multiple collections is an important consideration. They also suggest that progress to build a generic and robust ranking algorithm across the years of TREC *ad hoc* was not strong.

5. TREC GOV2 COLLECTION

In a second set of experiments, we studied a TREC web collection, Gov2, which consists of a crawl of US government web sites. There are two dominant types of documents present: HTML (92%) and PDF (8%); and it was noticed that amongst the Top Level Domains (TLDs) in the URLs crawled, two dominated: ".gov" (91%) and ".us" (9%).

As with the TREC *ad hoc* collections, a ranking of systems based on one sub-collection was compared to a ranking based on another. The correlation of the rankings was measured using τ, and significance was determined using a randomization test. Across the three years when Gov2 was used (2004-2006), a comparison between sub-collections split on document type was conducted, as was a comparison based on URL TLD. Six pairs of comparisons were possible (see Table 4).

For all six comparisons, it was found that τ was statistically significantly below what would be expected by random chance. All the comparisons were significant at the 0.001 level. The average τ measured between the six comparisons was 0.62.

Examining the differences between the HTML-PDF pairs, it can be seen that the average τ is 0.53, substantially lower than the range of the 1,000 random splits, whose lower bound is 0.7. There is a large difference between the ranking of retrieval systems based on the HTML and PDF documents of this collection. In other words, the system rankings on one part of this collection do not generalize to the other part of this collection.

The results also show that a ranking of runs measured on a web collection of ".gov" web pages are significantly different from the ranking obtained on ".us" web pages. This difference is not as strong as the difference found for HTML and PDF, however the effect is still present. The average τ for these three comparisons was 0.71, and is outside the bounds of the random distribution of τ values for each year.

6. REASONS FOR THE EFFECT

In the previous section, we demonstrated that using different sub-collections can lead to substantially different outcomes in terms of relative system performance as measured using standard IR evaluation metrics. As Table 3 shows, the particular pair of sub-collections that are used can have a large effect on the resulting level of Kendall's τ between system orderings.

TREC	Pair	τ	Random range from	 to	*p*-value
2004	HTML-PDF	0.47	0.84	0.99	< 0.000
	gov-us	0.79	0.87	0.98	< 0.000
2005	HTML-PDF	0.59	0.83	0.97	< 0.000
	gov-us	0.78	0.80	0.96	< 0.000
2006	HTML-PDF	0.52	0.70	0.92	< 0.000
	gov-us	0.57	0.57	0.89	0.001

Table 4: Testing significance using the randomization test.

The next stage of work was to determine which features of the sub-collections might influence the level of consistency (or inconsistency) that arises when they are used to evaluate retrieval systems. Intuitively, when two sub-collections have properties that somehow make them more "similar" to each other, then the expected agreement between evaluations would likely be higher than when the two sub-collections are less similar. We investigate three groups of such properties: the language used in the sub-collections; the length of document in the sub-collections; and, the distribution of relevant documents across the sub-collections.

In this section, we focus our analysis on the TREC 2-8 collections, since there are multiple sub-collections available under each collection. The Gov2 data does not readily lend itself to such analysis because the number of sub-collections is small.

6.1 Language similarity

At their core, most IR systems rank documents based on the distribution of terms within documents and across a collection. The language used in different sub-collections is therefore a feature that may indicate how these collections perform when searched by IR systems. We measure the similarity between two sub-collections, represented as a unigram language model, using the Jensen-Shannon Divergence (JSD), a symmetric measure of the similarity between two probability distributions [13]. A lower JSD indicates a smaller difference between the two language models, and therefore a higher level of similarity between the two sub-collections.

The relationship between the similarity of language used in each pair of sub-collections, and the correlation of run rankings using these sub-collections, can be quantified by calculating the Pearson correlation between the JSD and Kendall's τ for each pair of sub-collections. The results are shown in Figure 1. Across all seven TREC newswire collections, there is a substantial and statistically significant negative relationship (r=-0.509, p<0.0001). That is, when the language models of two sub-collections are more similar, then the relative system orderings obtained using those two sub-collections tend to agree more strongly.

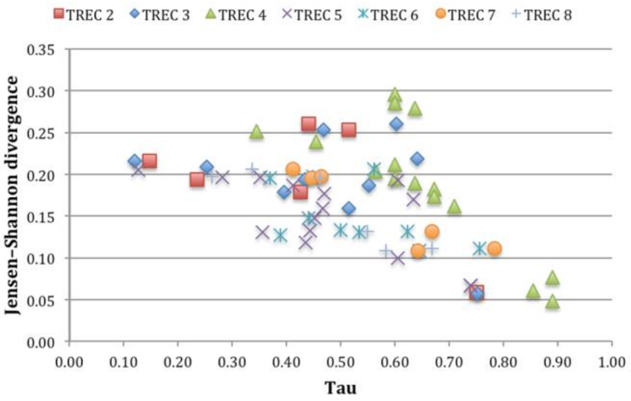

Figure 1: The relationship between the similarity of language used in sub-collections versus the consistency of system performance prediction on the same pair of sub-collections.

6.2 Document length

Most IR ranking algorithms include an explicit or implicit document length normalization component. It is therefore plausible that sub-collections that are more similar to each other in the size of documents would lead to more similar rankings of runs, and therefore to greater correlation of system orderings as measured with Kendall's τ. To investigate the relationship between average document lengths and system orderings, we first calculate the difference in mean document lengths for each pair of sub-collections. To enable comparison between different sub-collections, each difference was normalized into a range between 0 and 1. We then calculated the Pearson correlation between these relative document length values for each pair of sub-collections, and the τ value obtained from the ranking of runs for the same sub-collections.

The relationship between document length and system ordering agreement was negative, but not statistically significant (r=-0.172, p=0.161). That is, the difference in average document length between sub-collections did not affect how likely it is that system orderings based on a pair of sub-collections will agree.

We repeated the analysis, considering the length of only the *relevant* documents in each sub-collection, with the results shown in Figure 2. Here, the correlation was substantially stronger (r=-0.4075, p=0.0006). This suggests that when the difference in the average length of relevant documents between two sub-collections was smaller, then the system performance evaluations that are obtained using these sub-collections are likely to show stronger agreement.

6.3 Number of relevant documents

Typically, IR systems are evaluated using a performance metric, and these are ultimately based on the ability of a system to identify relevant documents. The distribution of relevant documents in sub-collections may therefore provide further insight into how likely it is that system evaluations on different sub-collections are likely to agree. For each pair of sub-collections, the difference in the available number of relevant

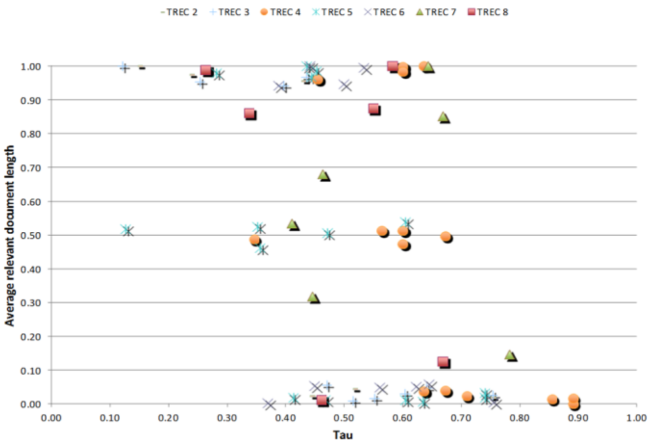

Figure 2: The relationship between the average lengths of relevant documents in pairs of sub-collections versus the agreement in relative system performance predictions on those sub-collections (as measured by Kendall's τ).

documents was calculated (to improve comparability the scores were again normalized into a range from 0 to 1). The Pearson correlation between the difference in relevant documents and τ values suggested an insignificant weak negative relationship (r=-0.090, p=0.464). This indicates that the prevalence of available relevant documents is not an important factor in determining the reliability of system evaluation across different sub-collections.

7. CONCLUSIONS AND FUTURE WORK

In IR evaluation research, it is often asked if inaccuracies of one kind or another in the evaluation process will significantly impede the measurement of effectiveness of retrieval systems. Almost always, the answer has been that although errors might affect the *absolute* number that one might measure for a particular system, if one is comparing IR systems then the error will affect both systems equally, and the *relative* ordering of the systems will almost always be unaffected.

This paper demonstrated that ranking systems using different sub-collections leads to substantial and statistically significant differences in the relative performance of retrieval systems. A number of alternate ways of forming these sub-collections were examined and in almost every case the formation methods resulted in sub-collections which caused erratic behavior in the IR systems. Despite years of working with such collections there does not appear to be any improvement in retrieval system consistency over such collections.

An initial analysis of possible causes of this behavior identified a significant correlation between the linguistic similarity of the sub-collections and the degree of erratic behavior in IR systems; the length of the relevant documents in the sub-collections had a similar relationship. However, differences in the number of relevant documents that are available in the sub-collections did not have a significant influence on the consistency of retrieval effectiveness experiments.

We view this paper as an initial step in the process of investigating sub-collections and their impact on retrieval. Although the properties of test collections have been examined in great detail in the past, it would appear that a study on the applicability of test collection results to other search situations needs further investigation. Examining other approaches to forming sub-collections, and studying sub-collections on a wider set of test collections and other types of search task, are part of the range of research that we intend to pursue.

The greatest promise from this research appears to be in understanding what it is about the IR systems that cause them to behave erratically on seemingly similar sets of documents. The work in this paper described differences in the effectiveness of different document ranking algorithms. What is not as yet understood is what part of the ranking algorithms is causing the differences. The term weighting schemes, document length normalization, language processing components, and query or document expansion methods, are all possible areas of investigation. With such an examination in place, the possibility of creating new ranking algorithms is enabled. Algorithms that detect sub-collections and adjust to deal with the properties of each part of an overall document collection have the potential to boost overall retrieval effectiveness. As document collections reach increasingly larger sizes, IR systems routinely partition the collections across the nodes of clusters. Examining whether improvements in retrieval effectiveness can be obtained while partitioning is an important goal.

8. REFERENCES

[1] C. W. Cleverdon, “The Evaluation of Systems Used in Information Retrieval (1958: Washington),” in *Proceedings of the International Conference on Scientific Information -- Two Volumes*, 1959, pp. 687–698.

[2] E. M. Voorhees, “Variations in relevance judgments and the measurement of retrieval effectiveness,” in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 315–323.

[3] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan, “Evaluation over thousands of queries,” in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 2008, pp. 651–658.

[4] J. Tague-Sutcliffe, “The pragmatics of information retrieval experimentation, revisited,” *Information Processing & Management*, vol. 28, no. 4, pp. 467–490, Jul. 1992.

[5] I. Soboroff, “Dynamic test collections: measuring search effectiveness on the live web,” in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006, pp. 276–283.

[6] E. M. Voorhees and D. K. Harman, *TREC: Experiment and Evaluation in Information Retrieval*, illustrated ed. The MIT Press, 2005.

[7] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, “Okapi at TREC-3,” *NIST SPECIAL PUBLICATION SP*, pp. 109–126, 1995.

[8] A. Singhal, C. Buckley, and M. Mitra, “Pivoted document length normalization,” in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, 1996, pp. 21–29.

[9] F. Scholer, A. Turpin, and M. Sanderson, “Quantifying test collection quality based on the consistency of relevance judgements,” in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information*, 2011, pp. 1063–1072.

[10] C. Clarke, N. Craswell, and I. Soboroff, “Overview of the TREC 2004 Terabyte Track,” in *Proceedings of TREC*, 2004, vol. 2004.

[11] E. M. Voorhees and C. Buckley, “The effect of topic set size on retrieval experiment error,” in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 2002, pp. 316–323.

[12] M. D. Smucker, J. Allan, and B. Carterette, “A comparison of statistical significance tests for information retrieval evaluation,” in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 2007, pp. 623–632.

[13] J. Lin, “Divergence measures based on the Shannon entropy,” *Information Theory, IEEE Transactions on*, vol. 37, no. 1, pp. 145–151, 1991.