Quantifying Test Collection Quality Based on the Consistency of Relevance Judgements

Falk Scholer
School of Computer Science &
Information Technology
RMIT University
Melbourne, Australia
falk.scholer@rmit.edu.au

Andrew Turpin
Dept of Computer Science &
Software Engineering
University of Melbourne
Melbourne, Australia
aturpin@unimelb.edu.au

Mark Sanderson
School of Computer Science &
Information Technology
RMIT University
Melbourne, Australia
mark.sanderson@rmit.edu.au

ABSTRACT

Relevance assessments are a key component for test collection-based evaluation of information retrieval systems. This paper reports on a feature of such collections that is used as a form of ground truth data to allow analysis of human assessment error. A wide range of test collections are retrospectively examined to determine how accurately assessors judge the relevance of documents. Our results demonstrate a high level of inconsistency across the collections studied. The level of irregularity is shown to vary across topics, with some showing a very high level of assessment error.

We investigate possible influences on the error, and demonstrate that inconsistency in judging increases with time. While the level of detail in a topic specification does not appear to influence the errors that assessors make, judgements are significantly affected by the decisions made on previously seen similar documents. Assessors also display an assessment inertia. Alternate approaches to generating relevance judgements appear to reduce errors. A further investigation of the way that retrieval systems are ranked using sets of relevance judgements produced early and late in the judgement process reveals a consistent influence measured across the majority of examined test collections.

We conclude that there is a clear value in examining, even inserting, ground truth data in test collections, and propose ways to help minimise the sources of inconsistency when creating future test collections.

Categories and Subject Descriptors

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

General Terms

Experimentation, Measurement, Performance

A preliminary version of this paper appeared at the 2010 Australasian Document Computing Symposium [14].

SIGIR'11, July 24–28, 2011, Beijing, China.
Copyright 2011 ACM 978-1-4503-0757-4/11/07 ...$10.00.

Keywords

Search engines, information retrieval evaluation

1. INTRODUCTION

The most common approach for evaluating an information retrieval system is through the use of a *test collection* using a standardized set of documents, topics, and human-generated relevance assessments, known as *qrels*. Test collections originate in the work of Thorne [17] and Gull [9] which inspired Cleverdon to create his Cranfield collections, defining the style of IR evaluation for decades [6]. Almost as soon as researchers described this evaluation approach, a number of criticisms were raised, many of which focused on the quality of the *qrels* and whether assessors would be able to reliably judge the relevance of documents to topics.

Over the nearly six decades that test collections have been used, many studies examined the impact of assessor mistakes on the quality of test collections. The format of these studies almost without exception used multiple assessors to calculate overlaps in judgements relative to each other. The studies showed that despite notable levels of disagreement, by and large, such noise did not affect the accuracy of test collections. These studies did not attempt to objectively calculate accuracy; neither did they try to measure an assessor's consistency across a set of judgements.

With the rise of crowd sourced relevance assessments, studies emerged measuring the accuracy of such judgments and, recently, there was work proposing models of assessor error with simulations of how such errors might impact on test collection based measurement. These recent studies highlighted that there has been no work to try to determine the levels of error made by traditional (and supposedly more trusted) test collection assessors; neither has there been an attempt to understand the nature of such errors. It is in this context that we conducted our research. The research questions that we investigate in this work are:

- Can one devise objective means of measuring assessor error?
- Are errors made by test collection assessors due simply to chance or are there discernible influences behind the mistakes?
- Are there alternate approaches to gathering relevance judgements that might reduce any measured error?

- Are there other consistent influences present in the relevance judgements of test collections that appear to affect the way that IR systems are evaluated?

The rest of this paper continues with a review of past work in this area. This is followed by a description of the data sets and methods used in our analysis. Next the results of a series of experiments on assessor error are presented. Experiments on other influences in relevance judgements are detailed after. Finally conclusions are drawn and future work is outlined.

2. RELATED WORK

Ever since the use of test collections became widely known, criticisms of the approach were described in the literature. In 1968, Katter [12] wrote that "a recurring finding from studies involving relevance judgments is that the inter- and intra-judge reliability of relevance judgments is not very high". Early test collection proponents, Cleverdon [7] and Lesk (with Salton) [13] both published studies comparing judgements by different assessors. They each concluded that despite a high level of disagreement between the assessors, the relative ranking of IR systems was largely unaffected by which set of judgements was used. These studies were repeated on a larger scale in a later study by Voorhees [19], in general drawing broadly similar conclusions, although Bailey et al. [2] showed that poor assessment can affect some aspects of measuring IR systems accurately.

The modern "standard" approach for gathering relevance assessments was established by TREC in the early 1990s [10]. A pooled set of documents are sorted by their document ID number (*DocID*) and judged in that order. The aim of such an approach is to minimise any influences on the assessors, who are unaware of how many retrieval systems returned a particular document or whether it was ranked high or low [10]. Other approaches to presenting pooled documents to assessors were proposed and their potential value discussed [24, 8]. When evaluated, the new methods were compared relative to the standard DocID sorting approach. However, these works did not consider the question of whether there is a way to determine if one approach is better than the other.

More recently, creating test collections using crowd sourcing (such as Mechanical Turk) grew in interest. Because workers in such approaches can be unreliable, there has been an increased focus on the potential of errors in relevance judgements. Alonso studied how best to obtain accurate judgements from workers using Mechanical Turk [1]. Carterette and Soboroff [4] suggested models of behavior in crowd sourced workers, simulating the impact of their posited patterns of mistakes on retrieval evaluation. A common approach to detecting errors by crowd sourced workers is to insert *ground truth* data (for which answers are already known) into the stream of items to be judged. If workers fail to mark up such data correctly, their inputs on other items can reasonably be assumed to be similarly mistaken. When building test collections, putting processes in place to check relevance assessors isn't common practice when non-crowd sourced assessors are used.

When researching the presence of duplicate documents in web collections, Bernstein and Zobel [3] pointed out that a number of duplicates in two web test collections (TREC's GOV1 and GOV2) had been "inconsistently classified" by relevance assessors. In effect, the duplicates were a form of ground truth data against which individual assessors could be tested. The extent to which assessors were examined in the paper was limited to calculating the fraction of inconsistently judged documents in the two collections studied.

The question of how to quantify assessor error has been considered in the context of e-discovery in the legal domain. The TREC Legal track interactive task simulates the process of reviewing documents in response to a request for production in civil litigation. In such a process, the relevance of a document (that is, whether the document should be considered to have a bearing on the case) is determined by a judge, called a *topic authority* in the TREC track [11]. It is therefore possible to objectively measure the rate at which relevance assessors make mistakes, by comparing their judgements against those of the topic authority [22]. However, despite a more objective notion of the "correctness" of relevance assessments, such comparisons are in essence still inter-assessor evaluations. Moreover, it is likely that the topic authority's own conception of relevance is subject to change over time.

From the review, it can be seen that examining the duplicates in test collections provides a means of determining the level of error made by relevance assessors, comparing different collections and different topics within collections. It potentially also enables comparisons to be made between different approaches for ordering *qrels*. As will be shown in the following section, the possible reasons for the inconsistent judgements made by assessors can also be explored.

3. METHODS AND DATA SETS

In this section we describe the test collections used and explain the methods that were applied to determine duplicate documents.

3.1 Test Collections

Collections from the Text REtrieval Conference (TREC) were chosen for our experiments for two reasons. First, the TREC approach to gathering relevance assessments is well documented and remained constant for many years. Second it has become a standard that many other collection formation methods are compared to. It was also realized by the authors of this paper that TREC files of *qrels* store the order that documents were shown to assessors. One can infer that the distance between two documents (for the same topic) in a TREC *qrels* file is proportional to the time between the two judgements being made.

This feature of the *qrels* was confirmed to us in a personal communication with Ellen Voorhees and Ian Soboroff at NIST, though with certain caveats. We are taking a simplified view of how distance equates to a difference in judgment time. Carterette and Soboroff [5] observed that more time is spent judging relevant documents than irrelevant documents. We experimented with counting relevant documents twice in the calculation of distance between two documents, assuming they take double the time to judge [5], but the results were consistent with simple counting, and so are not reported in detail. Another limitation of our assumption is that any breaks that may have been taken by assessors while judging documents for a particular topic are not reflected in the *qrels*. While it is theoretically possible that an assessor took an extended break from judging between every disagreeing pair of documents, this is highly unlikely, given the high number of disagreeing pairs per topic.

To analyse assessor behaviour, we examined relevance judgments over eight years of TREC, covering the early ad hoc test collections using mainly newswire and newspaper articles (TREC-1, 2, 3, 6, 7, 8) through to the more recent web collections (WT10G and GOV2). The ad hoc collections were grouped into three blocks, the first three years (labelled T123 in this paper), TREC-6 (T6, separated for reasons discussed later), and the final two years of TREC ad hoc (T78). Particular years were brought together if the collections used the same sets of documents.

3.2 Duplicate Documents

To identify pairs of highly similar documents that were both judged, a straight-forward document similarity approach was used. For each topic in a *qrels* file, all documents that have an entry (that is, an explicit relevance judgement) were indexed using the Zettair open-source search engine,[1] giving a *judged topic collection*. Next, each document with a judgement was submitted as a query to the previously created topic collection, with document similarities being calculated using the cosine measure [23]. The output of this step is a ranked list, ordering documents that have an explicit relevance judgment for a topic by decreasing similarity with the "query" document.

Manual inspection of duplicate documents from the collections showed that, almost without exception, documents retrieved with a similarity threshold of 0.9 or higher were very similar, near, or exact duplicate documents that a human assessor would be expected to judge consistently. Specifically, in the WT10G collection 58 duplicate documents were examined, 2 (∼3%) were found not to be duplicates. In T78, 400 randomly selected pairs of documents were examined producing similar level of incorrectly identified pairs. In the web collections (WT10G and GOV2), duplicate documents commonly arose due to mirrored web sites. In the news-based TREC collections, the duplicates were due to newswire sources that were either repeating the release of documents, or sending updates of stories as they developed. Because the 0.9 threshold was found to be a reliable identifier of duplicates, it was used throughout our experiments.

As will be seen in the next section, the number of duplicate documents present in the *qrels* of test collections varied substantially, which for some collections limited the number of tests that could be run. However, the aim of this work was to examine the feasibility of using duplicate documents as ground truth data. In future test collections, duplicates or near duplicate documents could be inserted into *qrels* to allow measurment of assessors.

4. RESULTS

The examination of the data covered a number of aspects, each of which is detailed in the following sub-sections.

4.1 Fraction of Inconsistently Judged Duplicates

We first measure the number of duplicates present in each of the collections studied, and compute the fraction of duplicates that were judged inconsistently by the assessors, as shown in Table 1. Only duplicates where at least one document was judged relevant are included in this analysis; that is, assessor consistency in judging non-relevant documents was not considered. The two web collections (WT10G and GOV2) have three levels of relevance judgements: highly relevant, partially relevant, and not relevant. Results for the trinary judgements, and for judgements converted to binary (folding the highly relevant and partially relevant judgements into one category) are shown separately. Two outlier topics were found in the collections: in T123, topic 056 had over 4,800 pairs of duplicate documents in its qrels; in GOV2, topic 775 had over 13,000 duplicate pairs. In both collections the statistics for those topics would have dominated the collection micro-averages, and were therefore removed.

Collection	Consistent	Inconsistent	(% of Total)
T123	689	120	15%
T6	228	45	16%
T78	308	71	19%
WT10G-binary	1413	293	17%
GOV2-binary	10138	2346	19%
WT10G-trinary	1374	332	19%
GOV2-trinary	9504	2980	24%

Table 1: Consistency of relevance judgments across TREC collections. The columns show a count of the number of pairs of duplicate documents that were judged consistently and inconsistently, followed by the percentage proportion of inconsistent pairs. Only duplicates where at least one document was judged to be relevant are included.

As can be seen, the total number of duplicates varied substantially across the collections, with earlier newswire collections (T123, T6, and T78) showing few duplicates in the *qrels*, WT10G an order of magnitude more, and GOV2, an order of magnitude more again. The fraction of duplicates that were judged inconsistently was relatively similar across the collections, although for the trinary judgements in the web collections, the number of inconsistencies increased.

Past studies on assessor consistency focused on comparisons between assessors (*inter-assessor* consistency). Using the duplicate ground truth approach shows that such comparisons were not just finding disagreements on the nature of relevance between assessors with, there was also a striking level of internal disagreement by the assessors of these test collections.

It is worth remembering that although the absolute numbers of errors in the second column of Table 1 are small, there is no reason to think that the measured inconsistency is limited to duplicates. Semantically similar documents in the *qrels* are just as likely to be at least as inconsistently judged.

Topic length

One reason for studying the early TREC collections, T123, was that the first 100 topics were particularly long, with 60.1 words per topic on average, compared to around 20-30 words per topic in later TREC collections. The fraction of inconsistently judged duplicates in the first 100 topics was examined and found to be slightly lower (13%) than the fractions for the other newswire collections. However the

[1] www.seg.rmit.edu.au/zettair

Judgement of Pair	WT10G	GOV2
0–1	271	2142
0–2	22	204
1–1	1297	8441
1–2	39	634
2–2	77	1063

Table 2: Number of duplicate documents judged consistently and inconsistently (highlighted in bold) across the different trinary judgement combinations of document pairs, measured in the two web collections.

difference wasn't judged large enough to be attributable to the longer topics.

While it may be that extremely long topics do reduce assessor inconsistency somewhat, there is insufficient evidence of such an effect for the data studied here.

Examining Disagreements on Graded Judgements

As noted above, the web collections WT10G and GOV2 were judged with three levels of relevance: not relevant (0), partially relevant (1), and highly relevant (2). We examined the numbers of each possible combination of document pairs in the two collections. As can be seen in Table 2 there is a strong commonality in the results. The number of pairs where assessors are consistent is largest for partially relevant, and lowest for highly relevant. Inconsistencies between not relevant and partially relevant are the most common, followed by partially relevant with highly relevant. The smallest number of pairs is for documents which assessors once thought weren't relevant and then later judged them highly relevant.

The frequency of inconsistent pairs is strongly skewed across the classes 0–1, 0–2 and 1–2: making a judgement call between non relevant and partially relevant documents contributed the great majority of inconsistent judgements, at 82% and 72% for the WT10G and GOV2 collections, respectively. Where trinary relevance assessments are being made in the construction of a test collection, the judgments assigned to partially relevant documents appear to require extra scrutiny.

Topic Effects

The two outlier topics (056 & 775) showed that there was variation in the number of duplicates across the test collection topics. For each topic, t, there is a set of consistent and inconsistent pairs, and for each pair, the fraction of inconsistent pairs was examined. For analysis to be carried out, a sufficient number of duplicates need to be present in the collections. We therefore use the WT10G and GOV2 collections; the other collections have fewer duplicate pairs than topics, making it unlikely that any meaningful analysis would be possible from examining these collections further.

To ensure that calculations were not affected by small numbers, only topics with at least 10 duplicate pairs were considered, and relevance judgements were converted to binary. To compute which topics had a fraction larger than might be normally expected, the data was compared to a random distribution of duplicate pairs across the topics. Any topics with a fraction of inconsistently judged pairs larger than three standard deviations from the mean of the randomly distributed fractions were noted.

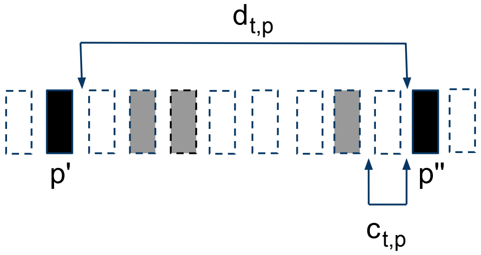

Figure 1: A hypothetical *qrels* file for a particular collection and topic. Each rectangle represents a document, in original *qrels* order. The two members of a pair of duplicates p are shaded in black and labeled p' and p''. The number of documents between these is $d_{t,p}$, 8 in this case. Intermediate documents that are highly similar to p' are shaded in light grey. The count of these, 3, is denoted by $s_{t,p}$ in the text. The number of documents between p'' and the closest similar document when moving back through the *qrels* is $c_{t,p}$, 1 in the example.

In the GOV2 collection, 26% of the topics that were measured had a significantly higher number of inconsistently judged documents than would be expected by chance; in WT10G the corresponding number was 25%. Across all topics in both collections, the fraction of document pairs that were inconsistently judged ranged between 36% and 79%, substantially higher than the overall percentages of 17% and 19% reported in Table 1. Again, it seems reasonable to assume that this level of assessor inconsistency was not just found in the duplicate pairs, but also in other documents judged for these topics. With such high levels of error, one might question the value of including such topics in a test collection. Equally, one might view such levels of error as a warning sign about the quality of outputs from the assessors who judged these topics.

4.2 Distance

Having established that there are a notable number of inconsistent judgements, we now consider the properties of the duplicates, starting with the distance between them. For each pair of documents, p, found for a topic, t, let $d_{t,p}$ be the total number of documents judged between the *first* and *second* of each pair (that is, the distance between the pair in the *qrels* file, see Figure 1). Remembering that *qrels* order can be reasonably interpreted as being proportional to the temporal order in which documents were presented, $d_{t,p}$ is used as a measure of time between the assessor judging the first and second documents of the pair.

We examined if the relevance assessors were more likely to judge duplicate documents inconsistently when those documents were seen further apart (that is, when there was a greater amount of time between seeing the documents). The average $d_{t,p}$ was computed for all p that were consistently judged, and then again for all p that were inconsistently judged. These values were then averaged across all topics in the test collections (that is, macro-averaged). The results of this calculation are shown in Table 3.

Collection	Inconsistent	Consistent	p-value
T123	79.3	63.6	insignificant
T6	133.3	32.0	p<0.01
T78	83.4	60.8	insignificant
WT10G	457.1	206.9	p<0.01
GOV2	213.0	131.2	p<0.01

Table 3: The average distance $d_{t,p}$ between inconsistently and consistently judged pairs of documents.

As can be seen, the distance between disagreeing pairs was always greater than the distance between agreeing pairs. Using a randomization test computed for each collection [16], it was found that the difference in distances was significant for every collection except T123 and T78. For the remaining test collections, the results show that the distance in the *qrels* between duplicate documents influenced the likelihood that they would be judged consistently. It would appear that assessors were not always clear on the criteria used to judge a document and that such criteria were either forgotten, or alternatively that an assessor's view of what constituted relevance shifted over time as the documents were judged.

4.3 Other Factors Affecting Judgements

Given that identical or near identical documents were being judged inconsistently at different times in the assessment process, we next examine the agreeing and disagreeing pairs in the GOV2 and WT10G collections in more detail, to identify factors that might influence a change in judgement.

Reminder Documents

It seems reasonable to assume that when an assessor encountered a duplicate of a previously judged document, but gave a different judgement than previously, that they forgot their previous judgement, and perhaps even forgot the document itself. Where other, similar documents are present between the two, the intervening documents could serve as "reminders" of the original duplicate, and thus lead to a consistent judgement of the second in a pair.

This indeed seems to be the case. For each topic, t, there is a set of consistent and inconsistent pairs, and for each pair, p, let $s_{t,p}$ be the number of documents that was *similar* to the first document, and occurred between the first and second in judgement order (see Figure 1). We continue to use $d_{t,p}$ to represent the total number of documents judged (the *distance*) between the first and second of each pair. Figure 2 shows the mean difference, for each topic, of the ratio $s_{t,p}/d_{t,p}$ for consistent pairs and inconsistent pairs. A high value indicates that there are more similar documents between consistent pairs than for inconsistent pairs for that topic. Fifty-nine topics from GOV2 and WT10G have a mean ratio greater than zero, and thirty-one have a ratio less than zero, with the mean difference being 0.091. It seems clear that having a high number of similar documents between a pair means that the pair is more likely to be consistently judged.

Given that the distance between similar documents that were judged inconsistently was greater than the distance between consistent pairs, perhaps this wasn't too surprising. Assuming that $s_{t,p}$ was small for most topics and pairs, and the previous section demonstrates that $d_{t,p}$ was larger for inconsistent pairs, then naturally the ratio $s_{t,p}/d_{t,p}$ will be smaller for such pairs. Define $c_{t,p}$ to be the number of documents between the second of pair p and its nearest similar document, in judgement order, towards the first document in pair p. That is, $c_{t,p}$ is the distance to the *closest* "reminder" document to the second document in p (see Figure 1). If the presence of reminders has the effect that it is more likely to lead to consistency between the first and second judgements of p, then $c_{t,p}/d_{t,p}$ should be smaller for consistent pairs than for inconsistent pairs.

Again, this intuition is supported. Figure 3 shows the mean over all pairs p for each topic t of $c_{t,p}/d_{t,p}$ for consistent pairs, minus the same for inconsistent pairs. Fifty-five topics have a similar document closer for consistent pairs than inconsistent pairs (a negative difference in the ratio) and thirty-three have the opposite. As the mean was less than zero (t-test, $p = 0.008$), we conclude that consistent pairs were more likely to have a document, close to the second, that was similar to the first of the pair, in judgement order. This close document is likely to have served as a reminder of the first document to the assessor, thus leading to the judgement of the second document agreeing with the first.

Inertia

A further possible factor that may affect relevance judgments is that assessors simply prefer to continue assigning the same value to documents. For example, Carterette and Soboroff [5] showed that autocorrelation exists in their judgement data from the TREC Million Query track: immediately after judging a document relevant, the probability of the next judgement also being relevant is significantly higher than the simple probability of a relevant judgement occurring independently of the previous judgement.

This *inertia effect* is present in the GOV2 and WT10G qrels, with the probability of a relevant judgement following a relevant judgement being 42% and 29% respectively, while the unconditional chance of judging a document relevant is 20% and 4%, respectively. Similarly, the conditional probability of judging a document irrelevant given that the previous was judged irrelevant is 86% and 97%, compared with the unconditional 80% and 96%. All four conditional probabilities are higher than the unconditional using a test on proportions ($p < 0.001$). These simple calculations from the qrels files, however, might reflect a clustering of documents in the collection, rather than a genuine bias in judging.

An alternative way to test the hypothesis that inertia exists can be carried out using the pairs of similar documents that were used in previous sections. In particular, if the second document in a similar pair was judged independently of the rest of the documents, it should have the same relevance score as the first document in the pair. We know that this is not the case for *inconsistent* pairs, however, so there must be factors other than document content influencing the judgement of the second document, one of which could be the judgement immediately preceding the judging of the second document.

Table 4 shows the probabilities involved, derived from counting data in the GOV2 collection. There are only a handful of disagreeing pairs for the WT10G collection, and while the trends are the same as those reported for GOV2, there is too little data to draw any firm conclusion from that collection. Clearly, the probability of judging the

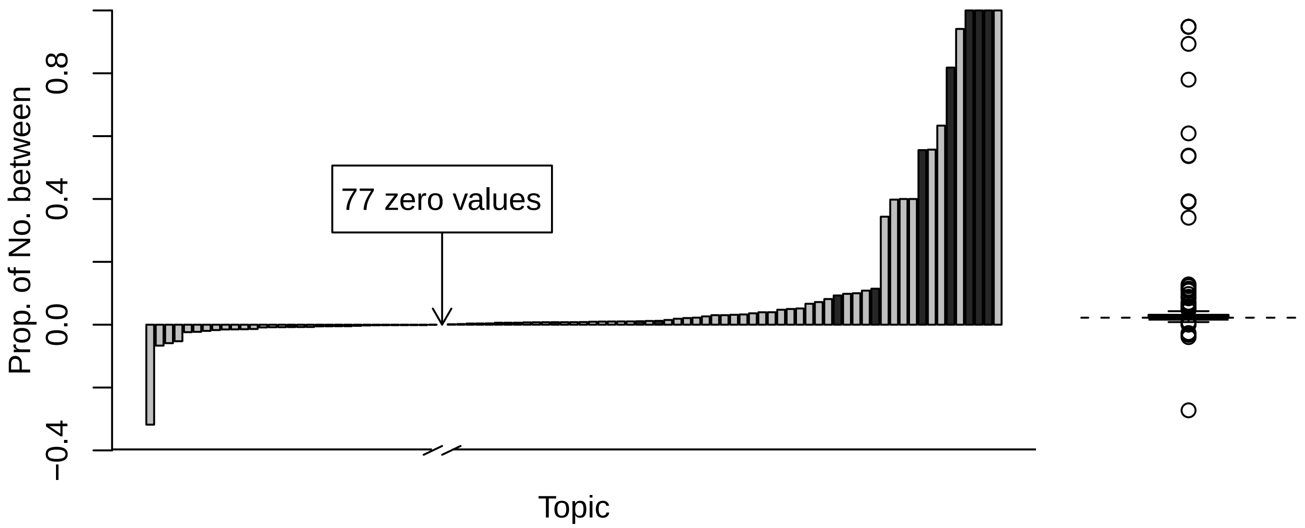

Figure 2: The difference between the mean number of documents similar to the first of a *consistent* pair that occurred between the first and second of the pair, as a proportion of the total number of documents between the pair, minus the corresponding quantity for *inconsistent* pairs. Topics that had no pairs with similar documents between were excluded (82 of 249 topics), and topics with a mean ratio of zero were not plotted in the barplot, but were included in the boxplot and statistics. Dark bars are topics from WT10G and light from GOV2. The boxplot on the right summarizes the values. The mean value is 0.058, which is different from zero (t-test, $p = 0.0003$).

Prob[Irrel.]	10.6%
Prob[Irrel. | preceding doc is Irrel.]	13.5%
	$p = 0.015$
Prob[Rel.]	3.0%
Prob[Rel. | preceding doc is Rel.]	8.0%
	$p < 0.001$

Table 4: Probability, derived from counting pairs of similar documents in the GOV2 collection, of judging the second document of a disagreeing pair either independently of the preceding document's judgement (rows 1 and 4), or conditional on the preceding judgement (rows 2 and 5). In both cases, the conditional probability was significantly higher (proportion test with p values as shown).

second document in a pair is conditioned on the previous judgement.

Testing Another Approach for Producing Qrels

In the TREC test collections, documents were presented to assessors in a consistent DocID sorted linear order. However, other orderings could be used; one example is an alternate set of *qrels* from Cormack et al. [8], whose work on an Interactive Search and Judge (ISJ) approach to assessment produced a set of *qrels* for the TREC-6 (T6) topics. Here, assessors searched for relevant documents for a particular topic by submitting multiple queries to an IR system. The authors of the paper kindly provided their ISJ *qrels* for us to examine. From this we computed the fraction of duplicate documents that were inconsistently judged and compared this to the numbers presented in Table 1. Note that the *qrels* produced by these two methods covered a different range of judged documents, which resulted in the different numbers of consistent and inconsistent judgements. As can be seen in Table 5, although the numbers are small, the fraction of inconsistently judged pairs is substantially smaller for the ISJ *qrels* than the DocID sorting method. A randomization test shows that the differences in percentages is statistically significant.

In the ISJ approach, the documents to be judged were presented to assessors as a series of ranked document lists. Since these are retrieved based on a similarity function, one would expect that duplicate documents occurred close to each other in such lists. This suggests that the time period between seeing duplicates was lower than in the linear presentation case, and so inconsistency rates were also likely to be lower. Similar, but non-duplicate, documents were also likely to be found close to each other in the rankings produced by the ISJ approach. From this result, we conclude that an ISJ approach to gathering *qrels* is potentially a more accurate approach compared to the DocID sorting method.

4.4 Correlations of System Performance

In addition to investigating the consistency of assessors at an individual judgment level, a further consideration is whether the demonstrated inconsistencies appear to impact significantly on system evaluation.

Computing the Tau of System Ranks

Measurement of an IR system using a test collection is in general only meaningful in the context of that collection,

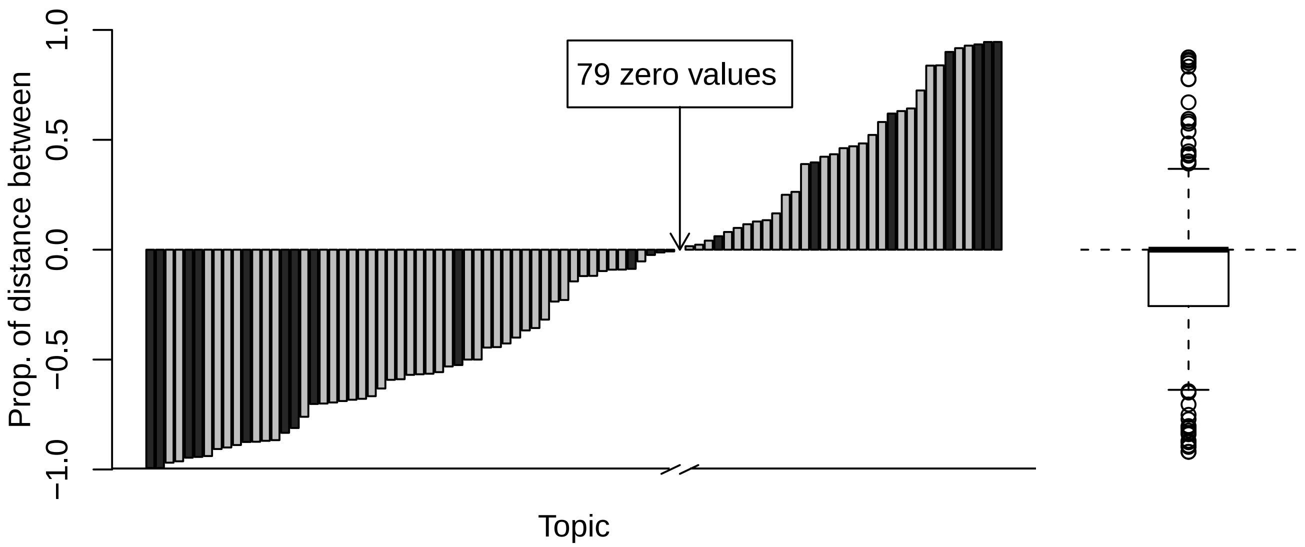

Figure 3: For a pair of duplicate documents, $c_{t,p}$ is the distance of the closest highly similar document that occurs between the pair in a *qrels* file. The figure plots the ratio of $c_{t,p}$ to the total distance between the duplicate documents for *consistent* pairs, minus the same ratio for *inconsistent* pairs. Details are as for Figure 2. The mean value is -0.09, which is different from zero (t-test, $p = 0.008$).

Collection	Consistent	Inconsistent	(% of Total)
T6-DocID	228	45	16%
T6-ISJ	417	35	8%

Table 5: The percentage of duplicate documents judged consistently and inconsistently. The final columns shows the number of inconsistent pairs as a proportion of the total number of pairs. Only duplicates where at least one document was judged relevant were considered.

and is not comparable for searches across different sets of documents [21]. For this reason, the focus of batch evaluation has been on *relative* system performance, rather than absolute. Where a set of queries has been run across a collection using a number of different search systems, the averaged effectiveness measures give an ordering of systems. When the same search systems are deployed on a different document collection, or with a different set of test queries, then even though the absolute scores are not directly comparable, a new relative ordering of system performance is obtained and these rankings can be compared.

A variety of measures are available to calculate such comparisons: Kendall's tau (τ) is commonly used in IR for this purpose. Tau measures the agreement between two ordered lists, and corresponds to the number of pairwise swaps that are needed to transform one ordering of one list to the other [15]. The value of tau is normalised to the range -1 $\leq \tau \leq +1$, with the extremes indicating perfect (inverse) agreement, and a value of zero indicating no association between the two lists.

Previous work on assessor consistency considered the level of agreement between system orderings when evaluation measures were calculated based on relevance judgements from different assessors. Voorhees [18] analysed this inter-assessor consistency by comparing system rankings based on *qrels* from different judges: TREC assessors and the ISJ approach of Cormack et al. [8]. She found that the average tau was 0.9, showing a relatively high level of agreement. This level is therefore representative of the kind of disagreement that might be expected when comparing *qrels* from different sets of people.

Given that it was shown that there is a DocID ordering to the *qrels* of many test collections, we focus on consistency across these ordered judgements sets. If there were systematic variations, they would warrant further investigation. We therefore proceeded as follows. First, the *ordered qrels* file was split in half, so that for each topic the first half of relevant documents were placed into an *early* partition, while relevant items that occur in the second half are placed into a *late* partition.

For retrieval systems, we used the official *runs* submitted to TREC for the ad hoc search tasks in TREC 2, 3, 6, 7 and 8; the Web tracks in TREC 9 and 10 (WT10G); and the Terabyte track in TREC 2006 (GOV2). Each set of runs represents a variety of different retrieval approaches, including both automatic (based only on a retrieval system) and manual (including human intervention) approaches. Because manual approaches often display strongly different characteristics, and in some cases reflect the ability of humans rather than the performance of search systems, we use only automatic runs in our analysis. Moreover, because submitted runs were from experimental systems, some submissions may include errors or unexpected behaviour; as a result, we follow standard practice and discard the bottom 25% of runs based on the MAP measure [20]. Finally, for the GOV2 collection, we note that the crawl of documents was carried out such that initially documents of type HTML were favoured, while later there was a bias

Collection	Tau	p-value	Overlap
TREC 2006	0.609	<0.001	0.667
TREC 10	0.628	0.047	0.333
TREC 9	0.702	0.273	0.538
TREC 8	0.396	<0.001	0.176
TREC 7	0.664	<0.001	0.667
TREC 6	0.466	<0.001	0.333
TREC 3	0.587	<0.001	0.429
TREC 2	0.456	<0.001	0.250

Table 6: Kendall's tau value between system orderings obtained when splitting the *ordered qrels*. The p-value indicates the result of a permutation test comparing the *ordered* and *random* split *qrels*. Overlap shows the change in the top 10 ranked systems when evaluated using the two halves of the *ordered qrels* (intersection divided by union).

towards PDF documents. To preclude possible interaction from this systematic variation in document types, we only used judgments for HTML documents in the analysis of this collection.

The level of Kendall's tau between system orderings based on the early and late partitions of the *qrels* files gives measures of agreement between system orderings, which can be compared to the 0.9 inter-assessor threshold.

We conducted a further comparison by partitioning the *qrels* into halves *randomly.* If there was a systematic change in relevance judgements that was related to judging order, then the tau score obtained by comparing the ordered partitioning would be lower than that obtained through random partitioning. In our experiments, we created 1000 random partitions of the *qrels*. These were used as a permutation test, to assess the significance of differences between the ordered and random levels of tau.

Results

The Kendall's tau correlation between system orderings based on the different partitions are shown for three TREC collections in Figure 4. For the GOV2 collection, shown in the left-most panel, the tau score for the *ordered* split qrels was 0.609, showing a relatively low level of agreement, especially in relation to earlier work on inter-assessor comparisons that resulted in a mean tau of 0.9 [18]. Moreover, the 1000 *random* split *qrels* led to tau values in the range from 0.764 to 0.946. Based on a permutation test, the intra-assessor effect is statistically significant ($p < 0.001$).

The tau scores, and p-values for a permutation test between the *ordered* and *random* split *qrels*, for the other TREC collections are shown in the first two columns of Table 6. As can be seen, results for the TREC 2, 3, 6, 7 and 8 collections were similar to those for GOV2: the tau scores between the two halves of the *ordered qrels* were substantially lower than the inter-assessor threshold of 0.9; and there was a statistically significant difference between the *ordered* and *random* split *qrels* ($p < 0.001$).

The only exception to this was the results for the TREC 9 and 10 Web track collections, as shown in the middle and right-hand panels of Figure 4. While still much lower than the 0.9 threshold, for TREC 9 the tau values for the *ordered* split were not significantly different from those obtained using *random* splits ($p = 0.273$) while for TREC 10 the difference was only slightly below the commonly used threshold for significance ($p = 0.047$). A possible explanation for this difference would be that the relevance judgments for the WT10G collection have a lower proportion of duplicate documents that are inconsistently judged; however, from Table 1 it can be seen that this is not the case. We also investigated whether there is a difference in the proportion of pairs that cross the "midpoint" at which the *ordered qrels* files are split for different collections, but again there appears to be no substantial variation.

Tau scores measure the agreement between two ranked orderings, taking the whole list into account. From a system evaluation perspective, however, we may be more interested in the top-ranked systems. It is therefore informative to consider the extent to which the set of the top 10 ranked systems changes when evaluation is carried out using the early or late halves of the *ordered qrels*. The right-most column of Table 6 reports the overlap between the two sets, measured as the intersection of the sets normalised by their union. For example, an overlap value of 0.333 indicates that 5 systems out of the top 10 are the same for both sets. The number of consistent systems in the top 10 sets ranges from 8 (overlap of 0.667 for TREC 2006 and TREC 7) to just 3 (overlap of 0.176, TREC 8). In particular, even the TREC 9 and 10 Web track collections show variation in the top 10 ranked systems (overlap of 0.538 and 0.333, respectively), despite showing less substantial variation in tau scores for the *ordered* versus *random* splits.

As can be seen from the analysis above, there appears to be substantial systematic variation in the way that retrieval systems are ranked between the early and later stages of the assessment process, leading to differences in the measured effectiveness of evaluation systems.

One possible source of such variation could be due to the inconsistency of relevance judgements, which are greater for judgements made further apart from each other. However, there are other plausible effects that are likely to also be contributing to the variation. In the earlier TREC test collections, DocID sorting of *qrels* has the unintended effect of grouping the documents by the news source they originated from. Therefore, the documents in the early partition will be from news sources, starting with letters early in the alphabet, such as the AP or Federal Register; documents in the later partition will more likely be from sources such as LA Times or Zipf. There should not be any influence of the source of an article on the relevance judgement or on the way that an IR system ranks that article. However, it would appear that such an effect is clearly present in the data shown here.

The order of documents in the *qrels* of the web collections appears to reflect the order in which the documents were crawled. The reason why an ordering effect exists in the GOV2 collection, beyond the document type effect discussed previously, is not as yet clear. We plan to investigate such factors in future work.

5. CONCLUSIONS AND FUTURE WORK

Relevance judgements are the key component for collection-based evaluation of IR systems. While there have been a number of studies examining the consistency of judgements made by different assessors, in this paper we examined the

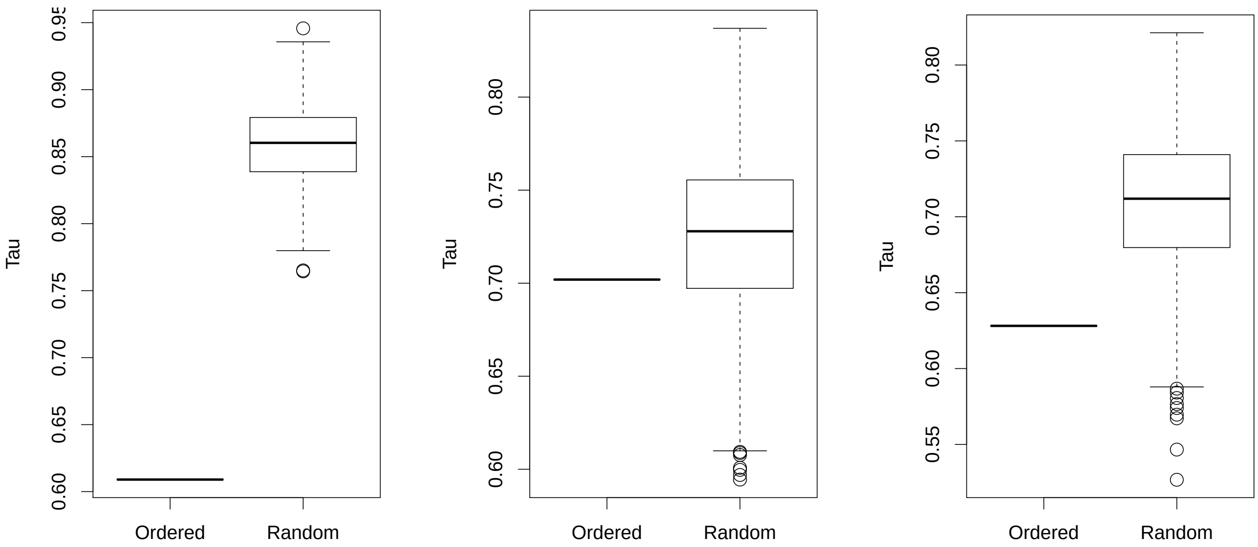

Figure 4: Kendall's tau correlation between system orderings when evaluated using the *ordered* and *random* split *qrels*. The three graphs show results for TREC 2006, TREC 9 and TREC 10, respectively. Corresponding statistics and data for the other TREC collections are presented in Table 6.

much less considered, but potentially more valuable, intra-assessor consistency.

Our first research question was how to objectively measure assessor error. By examining the inconsistent judging of duplicate documents in *qrels*, we demonstrated that one can measure a notable level of intra-assessor error in test collections. These error rates should be viewed as a lower bound, as it is expected that many more documents in the *qrels* will be subject to at least the same inconsistent assessment. Partially relevant documents were found to contribute disproportionately to assessor inconsistency. Means to identify problematic topics were described. Such a process can be used to guide test collection creation; for example, topics with high levels of inconsistency could be removed, or the work of the assessors judging these topics could be more carefully examined.

We next investigated whether the errors that were made by test collection assessors were simply due to chance, or if factors contribute to assessment mistakes. First, our analysis showed that the errors were impacted by the amount of time that passed between judgements being made on documents; it appears that assessors either forgot their criteria for judging the relevance of documents, or that their criteria changed over time. Second, assessors appear to have a judgement inertia: given a judgement on one document, assessors were predisposed to assign the next document with the same relevance value. A possible explanation for the inertia previously observed in qrels [5] was that relevant documents were clustered in the collection when viewed in *qrels* order, and so it is a feature of the collection, not the judges. However, our analysis in Section 4.3 showed that this was not the case in the GOV2 and WT10G collections; inertia existed for similar pairs that were not together in the collection. Third, judgements were shown to be influenced by the number of similar documents seen prior to judging a document. When similar documents were seen, these improved the consistency of judgments, serving as a reminder of the assessor's relevance criteria.

We investigated whether alternate approaches to gathering relevance judgements might reduce the measured error. An alternative set of *qrels*, built through an interactive searching and judging, process was obtained. The lower level of measured error in judging duplicates demonstrated that alternative methods for gathering judgments have the potential to significantly reduce the number of inconsistent judgements in a test collection.

Finally, we examined the impact of the ordering of *qrels* on the evaluation of IR systems. We calculated the correlation between system rankings obtained using judgements made early or late in the assessment process. The results indicated that there are significant, consistent and substantial differences in the way that retrieval systems are ranked when *qrels* are split based on the order in which those *qrels* are stored.

Although there have been a large number of studies on correlations between assessors, there has been little analysis of intra-assessor consistency. Neither have there been many examinations of problematic assessment on a per-topic basis. Overall, the results of this paper show that there can be great value in analysing the consistency of ground truth data. Our analysis and proposed diagnostic approach involved the examination of duplicate documents; however, in some collections, the occurrence of duplicates was a rare event. When building a test collection with few duplicates, one approach would be to manually insert duplicate documents at critical points of the judging process. We will examine such an approach in future work. We also plan to examine other types of duplicate or related documents that can be either found or inserted into collections. In particular, the data used in this paper is based on document-level relevance assessments. It would be illuminating to work with more fine-grained relevance data, where the specific sections of documents that are judged to be relevant are known. We will also examine the factors that are causing the low correlations in the system rankings from early and later splits in *qrels*.

It is rare for test collection creators to set up processes to check the quality of their assessors, or to try to identify poorly judged topics. Our analysis has shown that even in carefully run test collection exercises, notable levels of error are present. However, this simple analysis allows such errors to be identified and reduced. We hope that the practical outcomes of this work will be to encourage use of this analysis method, which can further enhance the quality of collection-based evaluation of IR systems.

6. ACKNOWLEDGMENTS

We thank Ian Soboroff and William Webber for helpful discussions, and Gordon Cormack and Charlie Clarke for providing the set of alternative *qrels* for the TREC 8 data. Andrew Turpin was supported by ARC Grant FT0991326.

7. REFERENCES

[1] O. Alonso and S. Mizzaro. Can we get rid of TREC assessors? Using mechanical turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 15–16, 2009.

[2] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the ACM SIGIR international conference on Research and development in information retrieval*, pages 667–674, 2008.

[3] Y. Bernstein and J. Zobel. Redundant documents and search effectiveness. In *Proceedings of the ACM CIKM international conference on Information and knowledge management*, pages 736–743, 2005.

[4] B. Carterette and I. Soboroff. The effect of assessor error on IR system evaluation. In *Proceedings of the ACM SIGIR international conference on Research and development in information retrieval*, pages 539–546, 2010.

[5] B. Carterette and I. Soboroff. The effect of assessor error on ir system evaluation. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 539–546, 2010.

[6] C. Cleverdon. The evaluation of systems used in information retrieval (1958: Washington). In *Proceedings of the International Conference on Scientific Information – Two Volumes*, pages 687–698. Washington : National Academy of Sciences, National Research Council, 1959.

[7] C. W. Cleverdon. The effect of variations in relevance assessments in comparative experimental tests of index languages. Cranfield Library Report 3, Cranfield Institute of Technology, 1970.

[8] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient construction of large test collections. In *Proceedings of the ACM SIGIR international conference on Research and development in information retrieval*, pages 282–289, 1998.

[9] C. D. Gull. Seven years of work on the organization of materials in the special library. *American Documentation*, 7(4):320–329, 1956.

[10] D. K. Harman. The TREC test collection. In E. M. Voorhees and D. K. Harman, editors, *TREC: experiment and evaluation in information retrieval*. MIT Press, 2005.

[11] B. Hedin, D. Oard, S. Tomlinson, and J. Baron. Overview of the TREC 2009 legal track. In *Proceedings of the 18th Text REtrieval Conference*, pages 4:1–40, 2010.

[12] R. V. Katter. The influence of scale form on relevance judgments. *Information Storage and Retrieval*, 4(1):1–11, 1968.

[13] M. Lesk and G. Salton. Relevance assessments and retrieval system evaluation. *Information storage and retrieval*, 4(4):343–359, 1968.

[14] M. Sanderson, F. Scholer, and A. Turpin. Relatively relevant: Assessor shift in document judgements. In *Proceedings of the Australasian Document Computing Symposium*, pages 60–67, Melbourne, Australia, 2010.

[15] D. Sheskin. *Handbook of parametric and nonparametric statistical procedures, 4th ed.* CRC Press, 2007.

[16] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the ACMCIKM international conference on information and knowledge management*, pages 623–632, 2007.

[17] R. Thorne. The efficiency of subject catalogues and the cost of information searches. *Journal of documentation*, 11:130–148, 1955.

[18] E. M. Voorhees. Variations in relevance judgements and the measurement of retrieval effectiveness. *Information Processing and Management*, 36(5):697–716, 2000.

[19] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36(5):697–716, 2000.

[20] E. M. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 316–323, Tampere, Finland, 2002.

[21] W. Webber, A. Moffat, and J. Zobel. Statistical power in retrieval experimentation. In *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management*, pages 571–580, Napa Valley, California, USA, 2008.

[22] W. Webber, D. Oard, F. Scholer, and B. Hedin. Assessor error in stratified evaluation. In *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management*, pages 571–580, Toronto, Canada, 2010.

[23] I. Witten, A. Moffat, and T. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishers, 2nd edition, 1999.

[24] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the ACM SIGIR international conference on Research and development in information retrieval*, pages 307–314, 1998.