Relatively Relevant: Assessor Shift in Document Judgements

Mark Sanderson *Falk Scholer*
School of Computer Science and Information Technology
RMIT University
Melbourne, Australia
{*mark.sanderson,falk.scholer*}*@rmit.edu.au*

Andrew Turpin
Department of Computer Science and Software Engineering
The University of Melbourne
Melbourne, Australia
aturpin@unimelb.edu.au

Abstract *The evaluation of information retrieval systems relies on relevance judgements – human assessments of whether a document is relevant to a specified search request. In the past, it was demonstrated that test collection assessors disagree with each other to some extent on the relevance of documents and can be inconsistent in themselves. This paper describes a series of investigations on assessor consistency, which demonstrate that the inconsistency of an assessor varies over time. We show that when documents are presented to assessors in a relevance independent order, documents judged as relevant appear to cluster. Examining pairs of documents in a sequence ordered by time-of-judgement, we find that relevance assessors judge highly similar document pairs more consistently when the pairs are seen soon after each other; the consistency reduces when the pairs are judged further apart. We contend that our analysis shows that changes are not due to random error, but instead reflect a* relevance shift*, whereby the assessor's conception of what constitutes a relevant document changes over time. Studying types of relevance judgement we find that the shift in judgements is greatest between highly and partially relevant documents. We also examine the impact of this inconsistency on how retrieval runs are ranked relative to each other and find that there appears to be a noticeable effect on such rankings.*

Keywords Information retrieval evaluation, Cranfield approach, Relevance judgements, TREC.

1 Introduction

Relevance judgements are a key component of test collections, which were defined in the "Cranfield methodology", the principal paradigm by which information retrieval systems are evaluated [5]. Soon after Cleverdon and colleagues proposed the use of test collections, objections were raised which focussed on the anticipated inconsistency of the assessors who would form such judgements. Such was the force of these criticisms that early IR test collection creators, Cleverdon and Salton, both conducted studies to understand the importance of assessor variability [5, 9]. They found that although assessors differed in their view about which documents were relevant, the way in which systems were ranked based on the different judgements was generally unaffected. As the size of test collections grew, the studies initiated by Salton and Cleverdon were repeated in subsequent decades coming to largely the same conclusions [21, 24].

While there has been a strong focus on examining the impact of assessor consistency, there has been less study on the nature of the inconsistency. Salton, Cleverdon and later Sanderson [10], showed that assessors tended to agree on the relevance of top ranked documents; they were less consistent on the lower ranked. Chen and Karger suggested that differences in assessment were linked to different interpretations of what a topic meant [4]. Bernstein and Zobel, examining the `gov2` collection, showed that individual assessors did make errors in judgement. They found a noticeable number of documents in a test collection that were textually almost identical but had been judged differently by the same assessor [1].

One of the potential sources of error in relevance judgements is the queries, which are often poorly specified. This can present a challenge to assessors on determining exactly what is and is not relevant. In the absence of a detailed specification, we hypothesise that assessors will look back at the documents they judged earlier to contrast with the document they currently have to assess. If assessors have a large number of documents to judge for a particular topic, earlier decisions may be forgotten and the documents used as a comparison to make relative judgements will change, leading to a

Proceedings of the 15th Australasian Document Computing Symposium, Melbourne, Australia, 10 December 2010.

shift in assessment criteria over time. In other areas where humans are used for assessment of documents, such as marking coursework, shifts in assessment are well understood; methods to control it such as detailed grade related criteria or even score standardisation tests are used to minimize this effect [12, 17].

To avoid biases in relevance assessments related to the rankings of retrieval systems, judges are presented with documents in the order in which they appear in the collection; in effect, an arbitrary order. Therefore, the data from the TREC collections that is stored in the relevance files (called *qrels*) is stored in the same order in which the documents were judged by the assessors (as explained by Harman [8], and confirmed in a personal communication with TREC staff). This data therefore provides a valuable resource from which it is possible to study relevance shift. This paper describes a series of preliminary experiments that were conducted on TREC data to better understand if relevance shift exists in test collections; if it does, what the nature of the shift is; and what impact any shift has on the way that retrieval systems are ranked by the test collection.

The rest of this paper is structured as follows: in Section 2 we present the background and related work on the evaluation of information retrieval systems, and previous studies into relevance assessments. Section 3 discusses relevance shift and how such a phenomenon might be identified in relevance judgements. We then present a series of experiments to test our hypotheses. Discussion and possible directions for future work are then given in Section 4.

2 Background and Related Work

Information retrieval systems are evaluated to determine how effectively they are able to help users fulfill information needs. The most widely-used approach for the evaluation of information retrieval systems is through the use of test collections. This approach, known as the Cranfield methodology, is a simulation of the search process [5]. A number of test *queries* are run across a fixed *collection* of documents using the retrieval system that is to be evaluated. For each query, the system generates a ranked answer list, with documents ordered by their estimated likelihood of being relevant to the search request. For each answer item that is returned, a human assessor then makes a *relevance judgement*, indicating whether the document is relevant to the search request, or not.

Based the answer lists of an IR system and the human relevance judgements, a range of performance metrics can be calculated to quantify the effectiveness of a retrieval system. These are generally based on precision (the number of relevant documents that were retrieved as a proportion of the total number of documents retrieved), recall (the number of relevant documents retrieved as a proportion of the total number of available relevant documents), or both. Mean average precision (MAP) is perhaps the most widely-reported performance metric. For a single query, average precision is defined as the mean of the precision scores obtained at each point where a relevant document is retrieved in a ranked answer list; MAP is then the mean of the average precision scores across a set of search topics. MAP has been shown to be a stable evaluation metric, and reflects both the precision and the recall of a retrieval system [2].

The Cranfield paradigm is used in IR evaluation campaigns including the Text REtrieval Conference (TREC) [22]. In TREC it is common practice for relevance judgements to be made by paid *assessors*, typically retired information analysts, who are asked to behave as if the provided search tasks are real information needs. The judging instructions stipulate that a document is to be judged as relevant if it contains any information that would be used in writing a report about the search topic under consideration. Answers are shown to assessors sorted by document number, to avoid potential bias from the ranking position of items in system answer lists; as a result, the judging instructions ask assessors to consider each document independently of all others (that is, there is no concept of redundant information). Moreover, to promote consistency of judgements, the documents to be judged for each topic are assigned to a single assessor [8].

Relevance is a vital concept for the evaluation of information retrieval systems, since it is the ability of such systems to provide useful answer documents – that is, documents that help a user to solve an information need – that determines their overall utility. While different levels of relevance have been proposed in the information science literature, in Cranfield-based evaluation of information retrieval systems it is typical to focus on *topical* relevance, where the focus is on the relation between a document and the topic under consideration [11]. In this operationalisation of relevance, user context is abstracted out, with the intention of allowing greater consistency of judgements. Nevertheless, many factors that can impact on the variability of relevance judgements have been identified, including requirements, statement variables, document variables, judgment conditions, judgment scales, and personal factors [6].

Despite limiting relevance assessments to be topical, analysis of judgements has shown surprisingly low levels of inter-rater agreement. A comparison of the relevance judgements of three different TREC assessors for TREC-4 topics showed an overlap of 0.42–0.49, while overlap with re-assessments of TREC-6 judgements by university students gave an overlap of only 0.33 [20]. In another study, re-assessment of 40 TREC newswire topics from TREC-6 to TREC-8 showed that around 64% of documents were judged differently [16]. However, despite these relatively low levels of agreement when assessing individual documents, relative *system orderings* obtained when competing IR systems are evaluated

based on different judgement sets were found to be generally stable, with a Kendall's tau correlation between system orderings of around 0.9 [20]. This level of correlation is widely taken to be representative as a threshold level of disagreement that should be expected when evaluating system orderings, due to noise from variation in relevance judgements.

While the design of IR evaluation campaigns seeks to limit inconsistencies in relevance assessments, it is widely acknowledged in the cognitive science community that people's choices and decisions are not consistent, and are sometimes not even rational [7, 13, 18]. In this paper we examine evidence for changes in the criteria applied by assessors when making relevance judgements.

3 Identifying Relevance Shift

To study shift in relevance criteria during the judging of a sequence of documents for a single topic, it is necessary to know the order in which relevance judgements were made by assessors. As described above, given a search topic t, the relevance data (*qrels*) are defined as a list of documents $\{d_1, ..., d_n\}$, where n is the number of documents judged for t. The assessors are shown the documents in the order 1 to n. While it is possible that assessors may stray from this order while judging, this form of behaviour is not thought to be common. We also specify R_d^t to be a *relevance judgement* detailing the relationship between t and d. For the TREC `wt10g` and `gov2` collections,

$$R_d^t = \begin{cases} 2, & \text{if } d \text{ is } \textit{highly relevant} \text{ for } t \\ 1, & \text{if } d \text{ is } \textit{relevant} \text{ for } t \\ 0, & \text{if } d \text{ is } \textit{not relevant} \text{ for } t. \end{cases}$$

For the TREC 7 and 8 *adhoc* collections only relevant and not relevant levels were specified (R_d^t is 0 or 1). Therefore, given a pair of documents d_i and d_j in the *qrels* list, we assume that the time period between the two documents being assessed is proportional to the *distance* $|i - j|$ between the documents in the list. With that assumption established, a number of tests were conducted to see if an assessor's view of relevance changed across the judgements. They are now described.

3.1 Distance between pairs of relevant and non-relevant documents

The first investigation attempted to check for clustering of relevant documents in the *qrels* list by comparing the distance between randomly selected pairs of relevant documents and similarly selected pairs of documents judged not relevant. If relevant documents are distributed uniformly throughout the *qrels*, then the distances should be equal, on average. If an assessor's conception of relevance shifted over time while judging the *qrels* list, then we would expect clustering of relevant documents in the *qrels* list. For example, similar

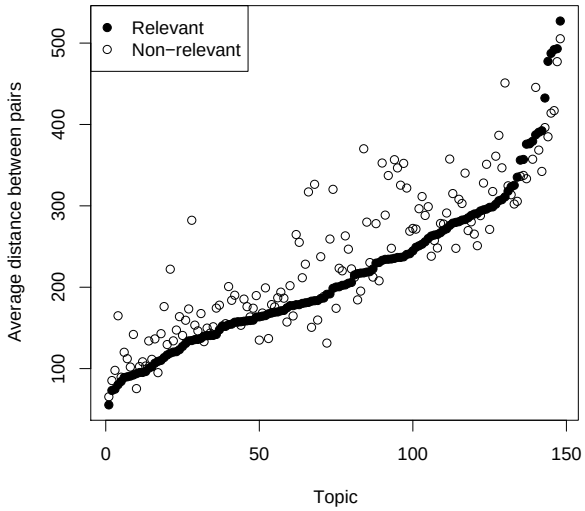

Figure 1: Average distance between pairs of relevant and non-relevant documents for the Terabyte track topics on the `gov2` collection. The x-axis has been sorted by increasing distance between relevant pairs.

documents found close to each other in the list may all be judged relevant, but an equally relevant document seen later may be judged irrelevant if the assessor shifts their relevance criteria.

It is important that other sources of relevance clustering are not present in the *qrels*, hence the TREC collections based on newspaper data could not be used. In such collections the document ids were related to the temporal order of the articles in the newspapers, so relevant documents would be expected to cluster around news stories reported intensely over a limited time period. It was decided to use *qrels* for topics 701 to 850 of the `gov2` web collection of ".gov" web pages. Here the document ids relate to the order in which the US government web sites were crawled.

This set of *qrels* still has some (known) features that may cause clustering of relevance judgements. First, due to the crawling process, documents from the same web site – which might be similarly relevant to a particular topic – were likely to be close to each other in the *qrels* list. Second, a large set of PDF documents were crawled towards the end of the gathering of the `gov2` collection, which means that most HTML documents are found in the first part of the *qrels* while most PDF documents are at the end.

Therefore, for each topic in `gov2`, we first examined the distance between randomly selected pairs of relevant documents that weren't from the same domain and were both an HTML file; second we tested for pairs that were both a PDF file. The result of this test is shown in Figure 1 where for each topic, the average distance between relevant and non-relevant pairs of HTML documents is shown. As can be seen, across a large number of topics the distance between relevant pairs of documents is smaller than the distance between non-relevant. Averaged across the topics,

	$R^t_{d_i} \neq R^t_{d_k}$	$R^t_{d_i} = R^t_{d_k}$
gov2	248	165
wt10g	455	207
t7t8	89	63

Table 1: Comparison of distance between pairs of similar documents in the *qrels* of three TREC test collections. Columns show inconsistent and consistent judgements respectively.

the distance was found to be statistically significant ($p < 0.01$). Here the significance test used was a randomization test, where multiple random partitions of the *qrels* were formed in the same per-topic proportion of relevant and non-relevant documents. When measuring distances between pairs of items randomly drawn from a set, the average distance is influenced by the size of the set. The distance between pairs drawn from small sets is likely to be shorter than for large sets. However, the difference in distance between the relevant and non-relevant was significantly larger than the difference found in the random sets. For more details on the randomization test's use in information retrieval, see Smucker [15].

The test was repeated for pairs of relevant and non-relevant documents where both were PDF files, and a similar statistically significant difference between the distances was found. Overall, the distance between randomly selected pairs of relevant documents in the gov2 *qrels* is smaller than the distance between similarly selected pairs of non-relevant documents. This implies that there are clusters of relevant documents in the *qrels*. However, it is possible that this is occurring because the relevance files were arranged in an order that is not independent of relevance. Therefore, the next experiment was conducted.

3.2 Distance between pairs of highly similar documents for which judgements are consistent or inconsistent

It has been noted in several past papers [1, 3] that within the TREC *qrels* sets, there are pairs of documents that are very similar to each other, but which assessors judged differently; there are also similar pairs that assessors judged consistently. Both types of document pairs occur in sufficient quantity for them to be studied in the context of this work. We hypothesised that there were two possible explanations for the inconsistent assessor behaviour: simple error, or relevance shift. If assessors were occasionally making mistakes, then the average distance between pairs that were judged the same or differently would be equal. On the other hand, if the inconsistency was due to relevance shift, then one would expect assessors to be consistent for pairs of documents seen soon after each other, and inconsistent for pairs seen far apart.

Here for a series of TREC test collections, the text of relevant documents from the *qrels* were in turn used as a query to search for other documents that were very similar to the "query document" and that had also been assessed for relevance. Similarity was calculated using the cosine measure across all content terms of judged documents (stemming and stopping were not applied). Any documents with a similarity of 0.9 or more were retained. Next, the distance in the *qrels* between the document pairs was measured. For this experiment, both web and newspaper based TREC collections were used. The concerns about clusters of relevant documents in the *qrels* of newspaper data was not a problem here, since if the inconsistent judgements were due to simple random assessor error there would be no difference in average distances.

	$R^t_{d_i} \neq R^t_{d_k}$	$R^t_{d_i} = R^t_{d_k}$
gov2	237	165
wt10g	510	207

Table 2: Comparison of distance between pairs of similar documents in the *qrels* of two TREC test collections. Columns show inconsistent and consistent judgements respectively. Here, only inconsistent pairs for *not relevant* and *partially relevant* are counted.

	$R^t_{d_i} \neq R^t_{d_k}$	$R^t_{d_i} = R^t_{d_k}$
gov2	265	165
wt10g	330	207

Table 3: Comparison of distance between pairs of similar documents in the *qrels* of two TREC test collections. Columns show inconsistent and consistent judgements respectively. Here, only inconsistent pairs for *partially relevant* and *highly relevant* are counted.

	$R^t_{d_i} \neq R^t_{d_k}$	$R^t_{d_i} = R^t_{d_k}$
gov2	236	165
wt10g	221	207

Table 4: Comparison of distance between pairs of similar documents in the *qrels* of two TREC test collections. Columns show inconsistent and consistent judgements respectively. Here, only inconsistent pairs for *not relevant* and *highly relevant* are counted.

The results are summarised in Table 1. As can be seen across all three tested collections, the mean distance between documents judged consistently (where the relevance of the document pair was the same) was substantially less than the distance between those judged inconsistently. We take this to indicate strong evidence that an assessor's view on what is and is not relevant changes over time.

As the wt10g and gov2 collections had two levels of relevance, there were three different types of inconsistent judgement:

qrels position	312	...	582	...	712
Doc	d_i	...	d_j	...	d_k

Figure 2: For a single topic, given a document d_i that is judged as relevant, d_j and d_k ($j < k$) are the furthest away documents that have cosine similarity with d_i of ≥ 0.9. The distance between d_j and d_k (130 in this instance) divided by the distance between d_i and d_k (400) is reported in Figure 3 for all relevant d_is.

- 0 and 1 - not relevant and partially relevant
- 1 and 2 - partially relevant and highly relevant
- 0 and 2 - not relevant and highly relevant

The differences for these three were tabulated in Tables 2, 3 and 4. The tables show that the distance between inconsistent judgements is smallest between the not relevant and highly relevant classes. This is particularly true for `wt10g`, where the difference in distance between consistent and inconsistent judgements is very small, indicating little or no relevance shift. For the `wt10g` collection the largest distance is between relevant and partially relevant documents, and for `gov2` it is between partially and highly relevant documents. In both cases, the more marginal relevant judgements appear to be more prone to shift than less marginal judgements.

3.2.1 Considering intervening documents

One possible reason why similar documents that are far apart in the *qrels* are judged inconsistently while closer similar documents are not is that there is less chance for the assessor to "forget" the relevance criteria used on the first document in the pair by the time they come to the second. If the two documents are close in the *qrels*, then only a small number of other documents have been judged, and so perhaps the assessor can maintain a consistent model of their relevance criteria. A further aid to maintaining a consistent relevance criteria could be that there are other documents between the pair that are also similar, and so they serve as a "reminder" to the judge of their criteria.

As a first attempt at measuring such an effect in the newspaper and web datasets we examined the similarity scores for documents between pairs, computing the distance of the closest document to the second of the pair. Specifically, for a pair of highly similar documents d_i and d_k, let d_j be the closest highly-similar document to d_k that lies between the pair in the *qrels*. Figure 2 shows a schematic of the approach we adopted.

For each relevant document d_i in the *qrels*, d_j and d_k are located, and the ratio of the distance between d_j and d_k to the distance between d_i and d_k is computed. The resulting triple is classed as *Same* if $R^t_{d_k} > 0$ (that is, both d_i and d_k are relevant), and *Different* if $R^t_{d_k} = 0$ (d_i is relevant while d_k is irrelevant).

`wt10g`	Same	1.7%	
	Different	1.5%	$p < 10^{-14}$
`gov2`	Same	2.7%	
	Different	1.3%	$p < 0.006$

Table 5: Proportion of documents in the interval between d_i and d_k in *qrels* order that have a cosine similarity score ≥ 0.9 to d_i. The final column shows the p value from a t-test of the row and the row above.

Figure 3 shows the ratios for the two collections. A ratio value of 1 indicates that there are no documents similar to d_i between d_i and d_k (that is, $d_j = d_i$), whereas a ratio close to 0 indicates that d_j was judged just prior to the judgement of d_k. Hence, we hypothesise that if the ratio is close to one, it is more likely the pair will be Different as there has been no similar "reminder" document before the judgment of d_k. It is clear in Figure 3 that the ratio is higher for Different pairs than for Same pairs in our two data sets. This difference is statistically significant (t-test, $p < 0.0001$ for `wt10g`, $p < 0.003$ for `gov2`), giving evidence that reminder documents are present for Same pairs, and not for Different pairs.

Also, simply counting the proportion of similar documents that occur in between d_i and d_k shows significantly more for Same pairs than for Different pairs, as shown in Table 5, again providing evidence that reminder documents may be present, and contribute to consistent relevance judgements, while a lack of such reminder documents leads to relevance shift.

3.3 Ordering retrieval systems based on subsets of relevance judgements

The criteria used to judge the relevance of documents may change over time, despite the best efforts of assessors. As we have argued, if this is the case, we would expect subtle differences between the documents that were judged as being relevant early in a series of relevance judgements, compared with those that are judged as being relevant later.

A further way to investigate such a shift is to consider the impact on performance scores when systems are evaluated using judgements from early in the *qrels*, compared with evaluating systems using judgements that were made later.

The effectiveness of information retrieval systems is commonly evaluated by calculating system performance measures such as MAP. Such scores are meaningful for a particular collection of documents, and set of test topics, but are not comparable across different collections or sets of search requests [23]. IR experiments therefore typically focus on *relative* effectiveness scores, for example, comparing the MAP scores of a "baseline" system with those of a "new" retrieval approach, with both systems running a common set of search topics over the same collection. Similarly, where a common set of queries is run over a single collec-

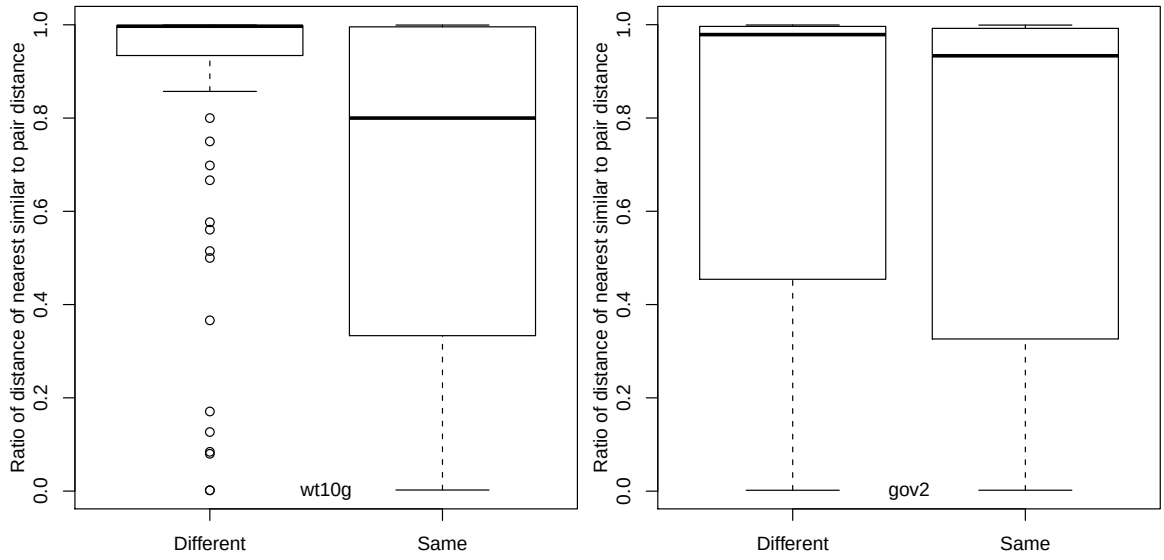

Figure 3: Ratio of the distance between the closest similar doc to the end of a pair (d_j and d_k) to the distance between the first and last of the pair (d_i and d_k) in *qrels* order.

tion with multiple systems, for example in evaluation campaigns such as TREC, it is the overall *ordering* of retrieval system performance that is of interest.

When a set of retrieval systems is evaluated using two different sets of relevance judgements, the relative performance of the systems may differ. The impact that changing relevance judgements has on the relative system scores can be measured using Kendall's τ (tau). Tau measures the extent to which two rankings agree, and is equivalent to the number of pairwise swaps that are required to obtain one ranking from the other [14]. The value of tau can range from +1 (perfect agreement) to -1 (perfect disagreement).

The assumption of relevance shift would suggest that systematic differences can arise between relevance judgements that were made at different stages of the judging process. In other words, we would expect there to be a difference in the relative ordering of systems when evaluated using relevance judgements from different parts of a *qrels* set (a tau score of less than 1). Moreover, if such a difference was systematic and reflected relevance shift, then the difference observed when using relevance judgements that were made early in the assessment process, compared to those that were made late in the process, should be greater than the difference that would be observed when randomly partitioning the set of relevance judgements (that is, not taking judgement order and possible relevance shift into account).

We test this hypothesis using data from the 2006 Terabyte track, which includes relevance judgements for topics 701–850 on the `gov2` collection. 80 runs were submitted to the track, representing different retrieval systems (or configurations of retrieval systems). Because runs that are submitted to TREC are based on experimental systems, they may contain bugs or have other problems. It is therefore common practice to discard the 25% lowest performing runs [21]. We discard the 20 runs with the lowest MAP scores based on the official full set of relevance judgements for our analysis below.

The original, ordered, relevance judgements are first partitioned into two halves, selecting the first 50% of relevant documents for each topic to be the "early" set, and the second 50% to be the "late" set. The 60 system runs are then evaluated based on their MAP scores when using the two partitioned relevance judgements. Kendall's tau between the two obtained system orderings is 0.493, showing substantial disagreement between the two orderings.

As a comparison, we randomly partition the relevant documents for each topic into two halves, and similarly evaluate the 60 runs using the split relevance judgements. Figure 4 shows the tau values obtained for 50 random partitionings of the relevance file. The mean tau score of the random runs is 0.752; moreover, the tau score from the split ordered relevance judgements is substantially lower than even the smallest value obtained for a random split, providing evidence for the presence of systematic variation in the ordered relevance judgements, and the impact that relevance shift may have on the evaluation of information retrieval systems.

4 Discussion

People's beliefs, opinions and criteria for making decisions change over time. In this paper we have carried out an initial set of experiments to investigate whether there is post-hoc evidence for the existence of relevance

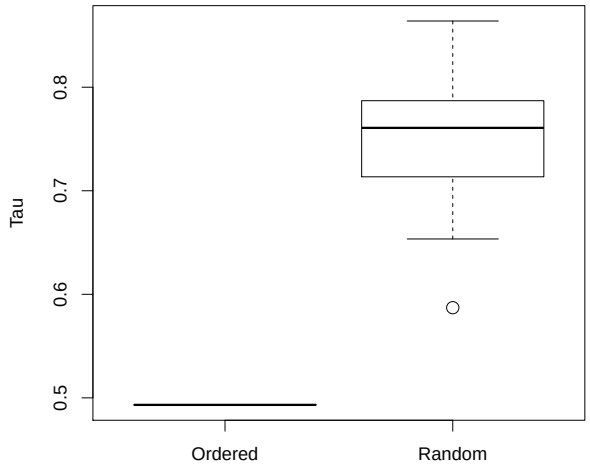

Figure 4: Kendall's tau correlation between system orderings when evaluated with split relevance judgements. using the original ordering versus random partitioning.

shift during judging of official relevance judgements made by TREC assessors.

Our results demonstrate that documents that are judged as being relevant tend to be clustered together as assessments are made, suggesting that coherent judging criteria are maintained most strongly in bursts.

One possible reason for this might be the mode in which TREC topics are developed and subsequently assessed. In particular, topics are chosen so that they will have relevant answer documents in the collection being used. Potential topics that do not have "enough" relevant documents are discarded, as are topics that have "too many" relevant documents [19]. Most times the assessment of relevance for topics is conducted by people involved in the topic development process, and so they have an a priori mental model of the number of relevant documents that should appear in the final *qrels*. During judging, then, if they have just deemed a run of documents to be relevant, they may alter their criteria to be more strict so that the total number of relevant documents does not get too high. Conversely, if there has been a run of non-relevant documents, they might loosen their criteria. Naturally any gross alterations in criteria could involve going back and re-judging documents subject to the new criteria, but perhaps this does not happen, or happens erroneously.

For pairs of highly similar documents, our results demonstrated that inconsistencies in judgements are not due to random error, but are again affected by the distance between the documents, with a greater distance leading to a higher likelihood of an inconsistent assessment. Furthermore, the presence of highly similar documents as assessments are being made impacts positively on the consistency of relevance judgements, suggesting that these help an assessor to maintain a consistent set of judging criteria. Finally, the presence of relevance shift appears to have a systematic impact on system assessments, with assessments based on judgements from early or late in a *qrels* showing much greater variation than when the judgements are partitioned randomly.

Taken as a whole, the results offer compelling evidence for the existence of relevance shift when large numbers of document judgements are being made for a search topic. However, our results so far have been from empirical analysis of relevance files, and have abstracted away from investigating the details of the documents themselves.

It should be remembered that in the experiments using pairs of highly similar documents, the presence of relevance shift is not restricted to just those pairs; relevance shift could well be occurring in judgements of documents elsewhere in the *qrels*. Therefore, in future work, we intend to carry out a user study where explicit assessments of pairs of documents from different parts of a *qrels* are made. We will also track which parts of documents are considered as being relevant, allowing more fine-grained analysis. User data will be key in further differentiating between cases where inconsistencies in judgements are made due to simple errors in attention, compared with shifts in relevance criteria. We also intend to further investigate the impact of the search topic specifications on relevance shift, using measures of clarity and readability to analyse the query statements.

Acknowledgements We thank Ian Soboroff for helpful discussions about the TREC assessment process and relevance judgements.

References

[1] Y. Bernstein and J. Zobel. Redundant documents and search effectiveness. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 736–743. ACM, 2005.

[2] Chris Buckley and Ellen M. Voorhees. Retrieval system evaluation. In Ellen M. Voorhees and Donna K. Harman (editors), *TREC: experiment and evaluation in information retrieval*. MIT Press, 2005.

[3] S. Bttcher, C. L. A. Clarke, P. C. K. Yeung and I. Soboroff. Reliable information retrieval evaluation with incomplete and biased judgements. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 63–70. ACM Press New York, NY, USA, 2007.

[4] H. Chen and D. R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 429–436. ACM New York, NY, USA, 2006.

[5] Cyril Cleverdon. The cranfield tests on index language devices. *Aslib Proceedings*, Volume 19, pages 173–192, 1967. (Reprinted in K. Sparck Jones and P. Willett, editors. *Readings in Information Retrieval*. Morgan Kaufmann Publishers Inc., 1997).

[6] Carlos Cuadra and Robert Katter. The relevance of relevance assessment. In *Proceedings of the American Documentation Institute*, Volume 4, pages 95–99, 1967.

[7] Benedetto de Martino, Dharshan Kumaran, Ben Seymour and Raymond J. Dolan. Frames, biases and rational decision-making in the human brain. *Science*, Volume 313, pages 684–687, 2006.

[8] Donna K. Harman. The TREC test collection. In Ellen M. Voorhees and Donna K. Harman (editors), *TREC: experiment and evaluation in information retrieval*. MIT Press, 2005.

[9] M. Lesk and G. Salton. Relevance assessments and retrieval system evaluation. *Information storage and retrieval*, Volume 4, Number 4, pages 343–359, 1968.

[10] M. Sanderson. Accurate user directed summarization from existing tools. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 45–51. ACM New York, NY, USA, 1998.

[11] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, Volume 58, Number 13, pages 1915–1933, 2007.

[12] M. N. K. Saunders and S. M. Davis. The use of assessment criteria to ensure consistency of marking: some implications for good practice. *Quality Assurance in Education*, Volume 6, Number 3, pages 162–171, 1998.

[13] James Shanteau. Psychological characteristics and strategies of expert decision makers. *Acta Psychologica*, Volume 68, Number 1-3, pages 203 – 215, 1988.

[14] David Sheskin. *Handbook of parametric and nonparametric statistical procedures*. CRC Press, 1997.

[15] 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*, pages 623–632. ACM New York, NY, USA, 2007.

[16] Eero Sormunen. Liberal relevance criteria of TREC – counting on negligible documents? In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 324–330, Tampere, Finland, 2002.

[17] I. Suto, R. Nádas and J. Bell. Who should mark what? a study of factors affecting marking accuracy in a biology examination. *Research Papers in Education*, pages 1–31, 2009.

[18] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, Volume 211, pages 453–458, 1981.

[19] E. M. Voorhees and D. Harman. Overview of the fifth text retrieval conference (TREC-5). In *The Fifth Text Retrieval Conference (TREC-5), Gaithersburg, MD, USA*, NIST Special Publication, Gaithersburg, MD, USA, 1996. Department of Commerce, National Institute of Standards and Technology.

[20] Ellen M. Voorhees. Variations in relevance judgements and the measurement of retrieval effectiveness. *Information Processing and Management*, Volume 36, Number 5, pages 697–716, 2000.

[21] Ellen M. Voorhees and Chris 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.

[22] Ellen M. Voorhees and Donna K. Harman. *TREC: experiment and evaluation in information retrieval*. MIT Press, 2005.

[23] William Webber, Alistair Moffat and Justin 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.

[24] Justin 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, Melbourne, Australia, 1998.