

Relevance Assessment: Are Judges Exchangeable and Does it Matter?

Peter Bailey*
Microsoft
Redmond, WA USA
pbailey@microsoft.com

Paul Thomas
CSIRO ICT Centre
Canberra, Australia
paul.thomas@csiro.au

Nick Craswell
Microsoft
Cambridge, UK
nickcr@microsoft.com

Arjen P. de Vries
CWI
Amsterdam, the Netherlands
arjen@acm.org

Ian Soboroff
NIST
Gaithersburg, MD USA
ian.soboroff@nist.gov

Emine Yilmaz
Microsoft Research
Cambridge, UK
eminey@microsoft.com

ABSTRACT

We investigate to what extent people making relevance judgements for a reusable IR test collection are exchangeable. We consider three classes of judge: “gold standard” judges, who are topic originators and are experts in a particular information seeking task; “silver standard” judges, who are task experts but did not create topics; and “bronze standard” judges, who are those who did not define topics and are not experts in the task.

Analysis shows low levels of agreement in relevance judgements between these three groups. We report on experiments to determine if this is sufficient to invalidate the use of a test collection for measuring system performance when relevance assessments have been created by silver standard or bronze standard judges. We find that both system scores and system rankings are subject to consistent but small differences across the three assessment sets. It appears that test collections are not completely robust to changes of judge when these judges vary widely in task and topic expertise. Bronze standard judges may not be able to substitute for topic and task experts, due to changes in the relative performance of assessed systems, and gold standard judges are preferred.

Categories and Subject Descriptors

H.3.4 [Systems and Software]: Performance evaluation (efficiency and effectiveness)

General Terms

Performance, Experimentation, Measurement

1. INTRODUCTION

Test collections for information retrieval (IR) typically consist of a corpus of documents, a set of topics, and a set of relevance judgements over a subset of documents from the corpus for each topic. Relevance judgements or assessments are made by people called

*Work carried out at the CSIRO ICT Centre.

Copyright 2007 Association for Computing Machinery. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of the U.S. Government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only. SIGIR '08, July 20–24, 2008, Singapore.

Copyright 2008 ACM 978-1-60558-164-4/08/07 ...\$5.00.

relevance judges or assessors (we use the term judges). The relative performance of IR systems can be compared by their ability to discover relevant documents over the set of topics using measures such as mean average precision (MAP) and normalised discounted cumulative gain (NDCG) [12]. In earlier times, it was possible to judge every document for every topic because the size of the corpus was so small. However, even by the mid 1970's, this became practically infeasible due to the increasing number of documents in the available corpora, and the pooling method was developed to select a likely subset of documents for review by the judges [13]. In the pooling method, unjudged documents are considered to be irrelevant, although it is known that relevant judgements may continue to be discovered from this set [4, 9].

TREC Enterprise 2007 developed a new test collection [3], with documents from Australia's Commonwealth Scientific and Industrial Research Organisation (CSIRO). Topics were created by the Organisation's science communicators and documents were judged for relevance by a variety of people inside and outside CSIRO.

As noted by Harter [11] and others, substantial variations in relevance judgements arise among different people, reflecting a host of possible backgrounds and experiences. We used the development of this test collection as an opportunity to investigate two issues. First, whether differences in topic and task expertise affect relevance judgements in any systematic way. Second, whether different relevance judgements consequently affect performance scores and rank ordering assigned to IR systems.

2. RELATED WORK

A number of studies have investigated whether variations in relevance assessment exist and how they affect measures of retrieval performance. These are summarised in Table 1. (A related body of work, investigating IR system evaluation measures and their correspondence with user satisfaction, is not the focus of this survey.)

Lesk and Salton's study [14] used eight people (a mixture of librarians and library science students), each authoring six topics. Each topic author judged their own six topics over the Ispra collection (1268 abstracts in the documentation and library science field), using a binary scale, and also judged six more topics from six different authors. All documents for all topics were judged by two people, one of whom was the topic author. Agreement levels averaged 0.31, using Jaccard similarity. The relevance assessments were used as input to three variations of the SMART system, which produced performance measures. Lesk and Salton found that despite the low level of agreement among judges, the recall-precision

researchers	relev. levels	topics	docs/ topic	judges/ topic	agreement	runs	performance	rank correl.
Lesk & Salton [14]	2	48	1268	2	31%, Jaccard	3	mean NP & NR	diff. in AP
Cleverdon [7]	5	32	200	4	—	19	NR	$r_s > 0.92$
Burgin [5]	3	100	1239	4	40–55%, Jaccard	6	AP	—
Voorhees & Harman [23]	2	49	400	2	72%, 3-way overlap	33	MAP	—
Voorhees [22] + Cormack et al [9]	2 + 3	49	≈ 124	2–5	33%, Jaccard	74	MAP	$\tau = 0.896$
Sormunen [19]	4	38	31–200	2	custom	—	—	—
Trotman et al [20]	2	15	67–135	3–5	custom	64	MAP	$r_s > 0.95$
this study	3	33	53–176	3	Cohen’s κ	15	infNDCG, infAP	τ

Table 1: Some characteristics of the empirical studies of inter-judge agreement and subsequent retrieval performance. Studies report a plethora of information; best efforts have been made to distill these here but additional data are available in the originals.

output measures remain “basically invariant for the collection being studied”. They explored four reasons why this may be so, and concluded that this is ultimately due to similarities in the documents retrieved early in the ranked lists by different methods, and considered relevant by either judge, as the performance measures used favour finding early relevant documents. Lesk and Salton discuss their work in terms of a strong and a weak hypothesis, both of which they found support for in their experiments. These can be readily understood in terms of changes to relevance judgements making changes to system scores (by some measure) or to relative orderings of systems (by their scores on some measure).

Cleverdon’s work [7] reports on re-judging documents from the Cranfield-II investigation by three additional judges, and used 42 topics, 19 index languages, and a five-level relevance scale. Rather than rejudging the entire corpus exhaustively, 200 documents were selected for each topic, using a non-random sampling method to focus on documents originally judged relevant, plus some additional documents from the corpus. Each judge re-judged all topics. The rankings of the normalised recall (NR) scores for the index languages were then compared using Spearman’s ρ for each of the three new judges in addition to the original Cranfield judges; correlations were found to be at least 0.92 for each combination of judges. Cleverdon concludes, similarly to Lesk and Salton, that performance measures and rank-orders remain similar despite differences in relevance assessments.

Burgin [5] reports on a study with a corpus of 1239 papers from Medline. Four groups of judges were involved: three groups were subject experts (one a professor of pediatrics and the topic author; one a set of nine faculty colleagues; and one a group of postdoctoral fellows in pediatrics) and one was a medical bibliographer. They carried out relevance judgements for 100 queries, using a three-level relevance scale. Mean inter-judge agreement levels using overlap agreement as per Lesk and Salton ranged from 40% to 55%. Burgin re-tested some of the explanations in Lesk and Salton’s study for why variations in relevance judgements did not materially affect the recall and precision results, in the context of his own study. He found support for their hypotheses, as well as three of their explanations for why these results held, despite noting differences in relevance scales, instructions to the judges, collection material, and nature of queries.

Harter [11] explored the past literature of empirical studies on how variations in relevance assessments affect measures of retrieval effectiveness. He noted: “*All find significant variations in relevance assessments among judges. And all conclude that these variations have no appreciable effect on measures of retrieval effectiveness, that is, in the comparative ranking of different systems.*”

He concluded by calling for an evaluation approach that reflects real world users; specifically trying to address and control for fac-

tors that lead to variations in relevance assessments. Our study pursues this approach with task and topic expertise.

In the TREC-5 overview [23], Voorhees and Harman report on inter-judge agreement experiments with a group of NIST-employed judges (retired information analysts), who also comprised the original topic authors. Assessment used a binary relevance scale, over 49 topics. The re-judged pools consisted of 400 documents per topic, with up to 200 judged relevant by the topic author and the balance made up of judged non-relevant documents. Each topic was re-judged by two judges. Harman and Voorhees found relatively high levels of three-way unanimous agreement (71.7%) compared to earlier studies. An observation was that judges agreed more on non-relevance than relevance. They also created sets of relevance judgements and compared system performance on MAP across the different judgement sets, reporting little apparent change in system rankings. The TREC-4 collection was many times the size of earlier collections, and also more diverse, but despite this the conclusions remained fundamentally the same.

This preliminary work was followed by a longer set of experiments by Voorhees comparing judgements from both the NIST judges for TREC-4, and NIST and University of Waterloo judges for a set of overlapped judgements on part of the TREC-6 dataset [22]. The Waterloo judges used an interactive searching and judging process to find and assess documents, using a three-level relevance scale. NIST judges used the standard TREC pooling method and a binary relevance scale. Correlation using Kendall’s τ was 0.896 for 76 systems ranked by MAP over the relevance judgements produced by NIST and the University of Waterloo. One topic (of 50) had no overlapping documents from both sets. This thorough analysis concluded that the relevance assessments created rarely have major effects on the relative system ordering. Voorhees adds caveats regarding situations when the number of relevant documents for topics is low, and when runs involve significant relevance feedback. Part of the explanation for this, she suggests, is instability in the measures used for performance in these circumstances. She also remarks on the minimum number of topics required to obtain acceptable correlation — her assessment was a minimum of 25 for this dataset. She concluded that the Cranfield approach and the TREC test collections were robust for their primary purpose — analysing the performance of retrieval algorithm variations when building an IR system. A twin of this study is the work reported by Cormack et al. in [9], although their focus is more on how to create test collections without the expense of pooling.

Sormunen’s study reported re-judging work comparing TREC’s binary relevance scale with a four-level relevance scale [19], but system performance measures were not the concern here.

Most recently, Trotman et al. report on a series of experiments for multiple judging of 15 INEX 2006 topics using between three

and five judges per topic [20, 21]. Rank correlation of the 64 systems was > 0.95 using Spearman's ρ , over a number of synthesised relevance judgement sets compared to the original baseline. Their conclusion was that exchanging judges makes little appreciable difference in rank order of systems, and suggest this would support an effective way to partition judging workload among individuals.

On a related issue, Mizzaro provides some frameworks for measuring disagreement between judges [15], but does not carry out empirical studies.

3. TREC ENTERPRISE 2007

The collection used in the TREC Enterprise 2007 track was created by crawling pages from CSIRO's public websites. These were distributed as the corpus to track participants, together with 50 topics. The topics were created with the involvement of science communicators from CSIRO. Their role is to communicate the science and business of CSIRO to the general public, including by publishing material on the public websites. The science communicators were asked to create topics for what was described as the "missing overview page" problem — where an overview page on a particular general topic was not currently available, and the science communicator was interested in creating one. This is a real job carried out by science communicators. Topics consisted of a short query (as might be issued to a web search engine); a longer description of the kinds of subjects that ought to be covered in the topic; a few key URLs for documents that already existed; and a few key contacts (human experts within CSIRO).

The track's document search task required each system to report a ranked list of documents relevant to the topic. At least one run from each participant had to be a query-only run without manual intervention of any kind. Individuals among the track participants carried out judging of pooled document results for each topic. Pooling was to depth 75 from the two runs per participating system given assessment priority by the participant. Judging instructions stated that the documents were to be classified on the basis of whether they would be good candidates to be included in a list of links in the new overview page. Documents were classified by each judge as "highly likely" to be included in this list; "possibly" included or useful; or "not" useful. This type of task-specific relevance classification is common to a large body of test collection creation activities in the Cranfield tradition, although the number of categories used may vary. Multiple judgement levels are more common in Web-oriented retrieval activities than classic ad hoc information retrieval evaluation.

Per-topic measures reported were average precision (AP) and NDCG with a gain of 2 for "highly likely" and 1 for "possibly" documents. To compute AP, only "highly likely" judgement classifications were considered relevant.

At a later time, a number of documents were re-judged by two sets of additional judges: as a "gold standard" documents were re-judged by the CSIRO science communicators who originally proposed the topic, and as a "silver standard" they were rejudged by science communicators from outside CSIRO. We consider "gold standard" judges to be experts in both the task and the topic. "Silver standard" judges are considered to have task expertise, but not topic expertise. Finally, the original "bronze standard" judges — TREC participants — have neither task nor topic expertise. Since the time of gold and silver standard judges was at a premium, we sampled documents from the pools to limit the number of judgements being carried out by these individuals.

Of the 33 topics for which we have complete overlap sample judgements (gold, silver, and bronze standard), a total of three gold standard judges, one silver standard judge, and nineteen bronze

standard judges were involved. (More silver judges participated, but unfortunately without overlap on topics from gold judges.) A total of 3004 documents were judged by these three sets of judges (compared to 22500 for the full pools for corresponding topics). There was an average of 91 documents per topic, with a low of 53 and high of 176.

4. EXPERIMENTS AND ANALYSES

CSIRO science communicators were the assumed system users for TREC Enterprise 2007 and represent the best possible judges. However, evaluations using the Cranfield model often substitute silver or bronze standard judges. Performance scores based on these lower-quality judgements do not measure performance for the assumed system user, and may diverge to the extent that silver or bronze standard judgements differ from the gold standard. Any such divergence could in turn lead to incorrect conclusions regarding a system's suitability for the task.

The questions we asked in the experiments here were:

1. How much agreement in relevance judgements is there between gold and silver standard judges, and between gold and bronze standard judges? A low level of agreement suggests there could be differences in performance measures.
2. If there is a low level of agreement, what effect does this have on performance measures (assumed to correlate with user satisfaction), and on TREC-style system orderings? Any difference in measures or system rankings would suggest that conclusions about performance are unreliable when based on silver or bronze standard judgements.

Measures for sampling methods were originally established using random selection of documents for judging [24]. This approach has recently been extended for non-random selection of documents [25]. Past work of Harman and Voorhees [23] suggested judges were more likely to agree that documents were irrelevant to a topic, than they were to agree on what constituted relevance. In situations where judging is limited or costly, it makes sense to focus judging effort on those documents more likely to be potentially relevant than those which are uncontentionally irrelevant. Indeed the pooling method typically used in TREC is exactly one such approach; also highly ranked relevant documents have more effect on performance measures than lowly ranked ones. As collections grow bigger, there arise more concerns regarding the creation of bias in the relevance assessments using pooling [4]. Investigations of non-random sampling methods in this context have also been conducted [18], which look to ensure sufficient samples are carried out in highly ranked documents due to their effect on performance measures.

In this section, we introduce infNDCG , a measure which like infAP [25] estimates a well-known performance score from a small number of judgements. A detailed derivation and analysis can be found elsewhere [25].

With infNDCG shown to be accurate, we examine the level of inter-judge agreement (Section 4.3) and the effect this has on system performance measures and rankings (Section 4.4).

4.1 infNDCG

Gold and silver standard relevance judgements were not available over the entire pool. We therefore use inferred measures to estimate what would have been recorded over full judgements. InfAP [24] is an estimator for average precision; infNDCG , which we describe here, estimates NDCG.

There are different versions of the NDCG metric depending on the discount function and the gains associated with relevance grades, etc. In this paper, we adopt the version of NDCG in `trec_eval`.

Let $r(1), (2), \dots, r(Z)$ be the relevance values associated with the Z documents retrieved by a search engine in response to a query q . Then, the NDCG value of this search engine can be computed as

$$NDCG = \frac{DCG}{N_q}, \text{ where}$$

$$DCG = \sum_{i=1}^Z r(i) / \log_2(i+1).$$

N_q is the normalisation constant for query q , chosen so that the NDCG value of a perfect list is 1.

Estimation of NDCG with incomplete judgements can be divided into two parts: (1) estimating N_q and (2) estimating DCG. Then, NDCG can be computed as $E[DCG]/E[N_q]$.

4.1.1 Estimating the normalisation constant (N_q)

The normalisation constant N_q for a query q can be defined as the maximum possible DCG value over that query. Hence, estimation of N_q can be defined as a two-step process: (1) For each relevance grade $r(j) > 0$, estimate the number of documents with that relevance grade. (2) Calculate the DCG value of an optimal list by assuming that in an optimal list the estimated number of documents would be sorted (in descending order) based on their relevance grades.

Suppose incomplete relevance judgements were created by dividing the complete pool into disjoint sets (strata) and randomly picking (sampling) some documents from each stratum to be judged. The sampling within each stratum is independent of the other, hence, the sampling percentage could be different for each stratum.¹

For each stratum s , let $nr_s(j)$ be the number of sampled documents with relevance grade $r(j)$ and let n_s be the total number of documents sampled from strata s and N_s be the total number of documents that fall in strata s . Since the n_s documents are sampled uniformly from strata s , the estimated number of documents with relevance grade $r(j)$ within this strata can be computed as

$$\hat{R}_s(j) = \frac{nr_s(j)}{n_s} \cdot N$$

Then, the expected number of documents with relevance grade $r(j)$ within the complete pool can be computed as

$$\hat{R}(j) = \sum_{\forall s} \hat{R}_s(j)$$

Once these estimates are obtained, one can use these values to estimate the value of the normalisation constant.

4.1.2 Estimating DCG

Given Z documents retrieved by a search engine, let $r(i)$ be the relevance grade of the document at rank i and $\log_2(i+1)$ is the discount factor associated with this rank. For each rank i , define a new variable $x(i)$, where $x(i) = Z \cdot r(i) / \log_2(i+1)$. Then, DCG can be written as the output of the following random experiment:

1. Pick a document at random from the output of the search engine, let the rank of this document be i .

¹Any sampling strategy could be thought of here. For example, an extreme case is where each document has a different probability of being sampled as discussed in Aslam et. al. [2]; each stratum can be thought of as containing only one document.

2. Output the value of $x(i)$.

It is easy to see that if we have the relevance judgements for all Z documents, the expected value of this random experiment is exactly equal to DCG.

Now consider estimating the outcome of this random experiment when relevance judgements are incomplete. Consider the first step of the random experiment, picking a document at random. Let Z_s be the number of documents in the output of the search engine that fall in strata s . When picking a document at random, with probability Z_s/Z , we pick a document from strata s .

Therefore, the expected value of the above random experiment can be written as:

$$E[DCG] = \sum_{\forall s} \frac{Z_s}{Z} \cdot E[x(i) | \text{document at rank } i \in s]$$

Now consider the second step of the random experiment, computing the expected value of $x(i)$ given that the document at rank i falls in strata s . Let $sampled_s$ be the set of sampled documents from strata s and n_s be the number of documents sampled from this strata. Since documents within strata S are uniformly sampled, the expected value of $x(i)$ can be computed as:

$$E[x(i) | \text{document at rank } i \in s] = \frac{1}{n_s} \sum_{\forall j \in sampled_s} x(j)$$

Once $E[N_q]$ and $E[DCG]$ are computed, $infNDCG$ can then be computed as $infNDCG = E[DCG]/E[N_q]$.

Note that this assumes that N_q and DCG are independent of each other, which may not always be the case. Better estimates of NDCG can be obtained by considering this dependence. For the sake of simplicity, throughout this paper, we will assume that these terms are independent.

4.2 Sampled judging

We were interested to investigate different methods for choosing documents for rejudging by gold and silver standard judges given at least one (bronze standard) labelling already. Two methods were used: topic-based sampling and effort-based sampling. These may be explained most easily with reference to an example.

Consider the judgements in Table 2 for two topics A and B. In topic-based sampling, we wish to apportion the sample size in accordance with the original pool size of each topic; for example, we might choose a 10% sample size. Thus we would select 10 documents from topic A and 20 documents from topic B. Placing emphasis on documents originally judged highly or possibly relevant, we might choose to select 60% of each topic's sample from the highly relevant labelled documents, 30% of the sample from the possibly relevant labelled documents, and 10% of the documents from the not relevant label. We would thus select documents in numbers as shown in the second part of the table. Note how this places re-judging emphasis on those documents originally labelled highly relevant — they form 60% of the sample to be re-judged, but constitute only 20% of the original pool. Note that the row sums for both topics are exactly 10% of their original size.

The effort-based method samples documents with some probability from each pool of labelled documents, irrespective of topic. The third row (labelled "total" in the table) is now the data we use for sampling. Using the same overall 60 : 30 : 10 ratio for the final 10% sample of 30 documents, we might end up with documents chosen from each topic as shown in the third part of the table. Note the column rows sum to the same numbers as for topic-based sampling, but the topic row totals do not.

<i>original judgements for 3 categories of relevance</i>				
topic	highly	possibly	not	total
A	10	20	70	100
B	50	50	100	200
total	60	70	170	300
<i>topic-based sample sizes, 60 : 30 : 10%</i>				
A	6	3	1	10
B	12	6	2	20
total	18	9	3	30
<i>effort-based sample sizes</i>				
A	10	1	2	12
B	8	8	1	17
total	18	9	3	30
<i>full sample sizes (topic+effort)</i>				
A	10	3	2	15
B	12	8	2	22
total	22	11	4	37

Table 2: Topic and effort sampling methods example.

In our experiments, we wished to support an investigation of which sampling method might prove better. With relatively little extra judging effort, this could be achieved by aggregating maximum numbers of documents per topic, as shown in the final part of the table. We refer to this as full sampling.

These three sampling methods (topic, effort, and full) were compared to the full measures evaluated over the bronze standard judgements (the only ones available with entire pool judgements). Selection percentages for the topic sample method were 60(*highly*) : 30(*possibly*) : 10(*not*), with replacements to make up the per-topic sample size coming in order from the more relevant sub-pools. (So for example, if there was only 1 document labelled “highly likely”, but the sample size said there should be 5, then an additional 4 documents would be chosen at random from the documents labelled “possibly” for the topic.) For the effort sample method, the percentages were 42 : 28 : 3 from the respective bronze standard labelled pools of documents (over all 50 topics).

The original definition of infAP [24] assumes that incomplete relevance judgements are a random subset of complete judgments. Throughout this paper, when we refer to infAP, we refer to the new version of the measure that can incorporate non-random judgements [25].

Calculations of infAP and infNDCG were compared to AP and NDCG (as computed by `trec_eval`; NDCG calculation is pre-official release) across all 33 topics and across all 15 automatic query-only runs. This approach (per-topic comparison) is intrinsically more likely to show differences, due to topic variation, than comparing scores averaged over all topics. Note that infAP and AP were calculated with “highly likely” and “possibly” judgements collapsed into a single “relevant” category for this purpose. (We also computed infAP and AP counting only the “highly likely” document judgements as relevant, but little difference exists between the two approaches.) Results are shown in Table 3. Kendall’s τ for both inferred measures is highly correlated with the corresponding full measures ($p < 0.01$) in all three methods. Error bars are calculated over 10 subsamples drawn at random from the full sample for the effort and topic sample methods. The effort sample method has

<i>Topic:</i>	AP		NDCG	
	τ	RMSe	τ	RMSe
topic	0.89±0.01	0.053±0.002	0.91±0.01	0.037±0.004
effort	0.86±0.01	0.062±0.008	0.80±0.02	0.063±0.007
full	0.92	0.037	0.92	0.029
<i>Mean:</i>	MAP		mean NDCG	
	τ	RMSe	τ	RMSe
topic	0.99±0.01	0.027±0.003	0.97±0.02	0.005±0.002
effort	0.96±0.02	0.012±0.007	0.92±0.02	0.011±0.002
full	0.98	0.020	0.98	0.003

Table 3: Kendall’s tau correlation (τ) and Root Mean Squared error (RMSe) for original vs inferred measures. For topic and effort sampling, figures are mean \pm one standard error.

the lowest correlation, possibly because the effort sample method was based on pools formed from the labels over all 50 topics, rather than the final 33.

Kendall’s τ provides us with an understanding of how well the different measures track the original, but there could still be a large difference in absolute scores. To assess how closely the scores are related, we use root mean square error (RMSe in the table). The worst RMS error value is 0.063 ± 0.007 (using infNDCG and effort sampling), and other combinations are lower than this.

The measures of correlation and RMS errors originally reported [24] are based on scores averaged over all topics. We show the results for averaged scores in the lower half of Table 3. The correlation and error rates compare favourably with the earlier work.

These results demonstrate that the two inferred measures appear to be reliable estimators, and can be calculated with much less judging effort. In the remainder of this paper, we report infAP and infNDCG over 33 topics as a measure of performance using the full sample (since, probably because it has more documents in the sample, it has the highest τ and lowest RMSe).

4.3 Inter-judge agreement

InfNDCG and infAP are reliable predictors of NDCG and AP, but will differ as relevance judgements disagree. We therefore considered the level of agreement between judges over the TREC Enterprise documents and topics.

Table 4 summarises the level of agreement between judges, as conditioned probability distributions. For example, of those documents classified as not relevant by bronze standard judges, 5% were classified as highly relevant by gold standard judges; of those documents judged highly relevant by bronze standard judges, 36% were classified the same way by silver standard judges.

Several trends are apparent in this simple analysis. First, there is broad agreement between all groups of judges on irrelevant documents, with between 81% and 89% agreement between each pair. There is also reasonable, although lower, agreement on which documents are highly likely to be included in the putative overview page. The lowest agreement in this case has 36% of documents marked highly by bronze standard judges marked the same way by silver standard judges; these are the two sets of judges without topic expertise. There is most disagreement on documents marked as possible, although the agreement still appears better than chance alone.

Overall, silver standard judges rate fewer documents in the top two categories than do gold standard judges, which suggests scores

	gold judgement		
	highly	poss.	not
bronze highly	0.58	0.18	0.24
bronze poss.	0.35	0.25	0.40
bronze not	0.05	0.13	0.82
gold overall	0.39	0.19	0.42

	silver judgement		
	highly	poss.	not
bronze highly	0.36	0.28	0.36
bronze poss.	0.17	0.30	0.53
bronze not	0.02	0.09	0.89
gold highly	0.46	0.32	0.22
gold poss.	0.13	0.31	0.56
gold not	0.05	0.14	0.81
silver overall	0.23	0.24	0.53

Table 4: Agreement between judges (probability distribution of gold or silver standard judgements, given bronze or gold standard judgements). See text for details.

based on these judgements would be lower. Bronze standard judges are more willing than their gold standard counterparts to mark documents as possibly or highly likely to be included: there is only 58% agreement with the gold standard in the “highly likely” category, and 24% of the documents marked highly by bronze judges were considered irrelevant by the gold standard. We may expect higher scores from these judgements than from the gold standard.

Further analysis used Cohen’s κ [8], which is a chance-corrected measure of the agreement between two judges classifying objects into arbitrary categories. Negative values indicate less agreement than would be expected by chance alone, zero indicates agreement due only to chance, and higher values (up to 1) indicate agreement beyond that expected by chance. In the field of computational linguistics, Carletta made a case for κ in inter-coder agreement [6], and it has become a standard measure there, albeit with ongoing discussions about the appropriate way to interpret what values signify degrees of agreement [10]. A recent survey article by Artstein and Poesio [1] provides a thorough review of the many considerations in the use of the κ measure, its assumptions, variants, and alternatives. Here we use a variant due to Siegel and Castellan [17], and following Rietveld and van Hout [16] consider $\kappa \geq 0.6$ to indicate substantial agreement.

Agreement between gold and silver standard judges, calculated over the labelling of all sampled documents for each topic, is illustrated in Figure 1(a). Although in 25 of the 33 topics agreement is significant ($p < 0.05$), in only 3 topics is it substantial ($\kappa \geq 0.6$). Similar results were observed for the gold and bronze standard judges (Figure 1(b)): significant but small agreement in 22 of 33 topics and significant and substantial agreement in only one. Regardless of the debate over the measure, these results fail to indicate much agreement exists between judges.

Note that due to the sampling method used (Section 4.2), most documents seen by gold or silver standard judges had already been marked as “highly likely” or “possibly” by a bronze standard judge. If all judges classified a uniform random sample drawn from the full set of documents in the corpus, we may expect higher levels of agreement reported by κ , because the expectation of seeing a

relevant document is low in such a sample, and it is easy for judges to detect obviously irrelevant material. In general, however, we are less interested in these “easy” cases where documents are clearly irrelevant; and disagreements about these documents are less likely to impact performance scores for a well-performing IR system.

This low level of agreement between judges, over a substantial majority of topics, suggests that measures including infAP and infNDCG may vary substantially if judges are exchanged. Further analysis investigated this possibility.

4.4 System performance

TREC runs vary in the amount of information they use and the way they process the query. In the following analyses we have considered the fifteen runs which were reported as fully automatic, using short query titles only, as they are most representative of current search systems.

This selection of runs avoids having many runs from the same system, which often vary in small and subtle ways due to algorithmic experiments being carried out by participants. Past work has indicated that systems differ more between each other, than from variations in a single system. We thus chose not to include such variants since small differences in judgements may lead to perturbations in system ordering (and falsely drawn observations that changes in ordering occur) that are not good indicators for more significant differences between systems.

4.4.1 Effect on performance

A simple analysis considered the effect of variations in judgements on each system’s performance. Inferred scores, both infAP and infNDCG, were calculated for each selected run and topic for all three sets of relevance judgements. We take the score at each topic as an indication of each run’s performance, and use the Wilcoxon matched-pairs signed-ranks procedure to compare performance according to each set of judgements.

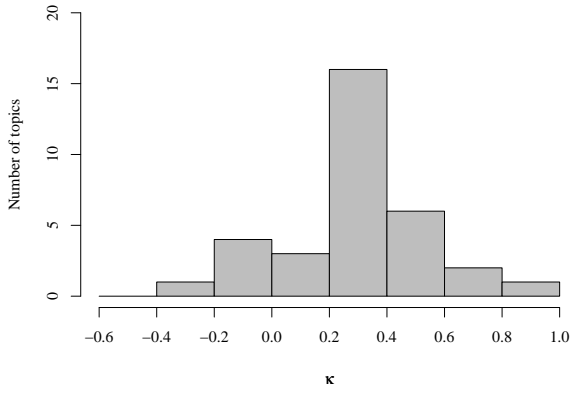
There was a consistent, but small, difference in infAP scores calculated from gold or silver standard relevance judgements (Figure 2(a)). Of the fifteen runs considered, eleven performed significantly worse on silver standard than gold standard judgements, with absolute differences in mean infAP of 0.01 to 0.07 ($p < 0.05$). Performance appeared higher using bronze standard judgements than gold standard judgements; six of fifteen runs had absolute improvements of 0.08 to 0.10 ($p < 0.05$). This is consistent with our observations in Section 4.3.

Differences were less consistent when systems were scored by infNDCG (Figure 2(b)). Only one run was significantly different with silver standard judgements (a drop of 0.02), and five runs were significantly better with bronze judgements (a gain of 0.06 to 0.08). Again, this is in agreement with earlier observations.

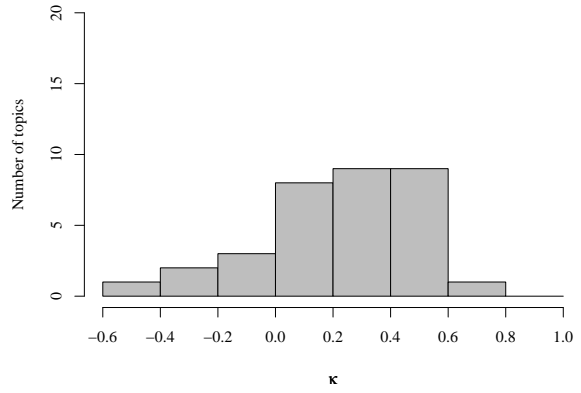
Overall, infAP is more likely to differ between judges than is infNDCG. Bronze standard judges also show more variation from the gold standard than do silver standard judges. The absolute differences in mean scores are small, however, even in those cases where they are statistically significant.

4.4.2 Effect on system ordering

Ranked lists of runs, ordered by performance scores, are commonly reported and can be used to identify the best-performing systems. If changes in relevance judgements, and hence on performance scores, were to perturb these ranked lists then medium- or low-performing systems may mistakenly be identified as high-performing; this could influence future development and possibly purchasing decisions. A final analysis therefore investigated the difference in these ranked lists as judges are changed.

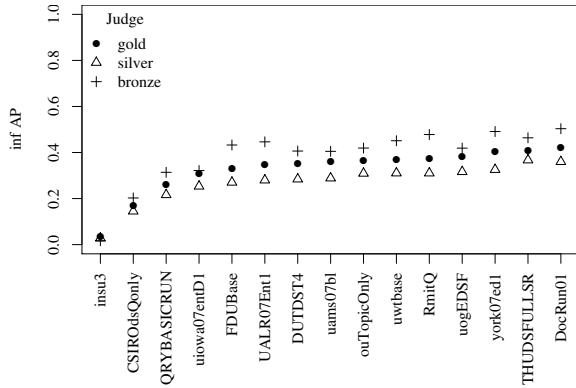


(a) Gold and silver standard judges.

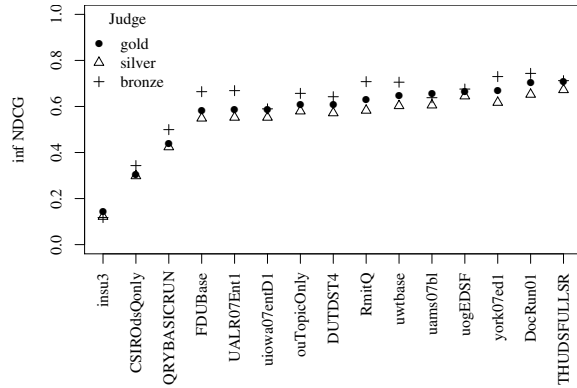


(b) Gold and bronze standard judges.

Figure 1: Agreement between judges (Siegel and Castellan’s κ). Total 33 topics.



(a) infAP



(b) infNDCG

Figure 2: Mean infAP and infNDCG scores, over all 33 topics, for each selected run and for each judge.

When systems were ranked by mean infAP, there was little difference in order whether gold or silver standard judgements were used ($\tau = 0.96$). The same was true for gold and silver standard judgements when systems were ordered by mean infNDCG ($\tau = 0.94$). Comparing orderings produced by gold and bronze standard judgements, however, showed larger differences: $\tau = 0.73$ and 0.66 for mean infAP and mean infNDCG respectively.

Some consistency in determining the top performing systems was observed, more noticeably with gold and silver judgements. The top four runs were ranked highest on both mean infAP and mean infNDCG (though not in the same order) by these judges. Bronze judgements agree in identifying the top three performing systems with the gold standard using infNDCG. All agree on the three lowest performing systems with both measures.

It seems therefore that the document-level disagreements observed between gold and silver standard judges do not have a material effect on system rankings. Rankings based on bronze standard judgements however show greater difference, and where judges of this type are used rankings should be treated as approximate only.

5. DISCUSSION

Our experiments have some limitations. First, although our 33 topics are more than the 25 recommended for stability by Voorhees [22], more topics would be better. Similarly, having more judges involved in each category would increase confidence that effects being seen were due to the task and topic expertise level of the

judge. Second, the results we have found may be due to the particular task and organisation considered here, and evaluation of different tasks or organisations may not show similarly consistent differences. Third, it would have been interesting to control the sampled judging process with additional bronze judges doing re-judging. Finally, we did not control for other factors which may affect judgements, beyond the grouping of judges into expertise categories and presentation of documents in random order.

Future work will increase the number of topics and judges involved. We also intend carrying out further analysis to investigate if there are additional systematic sources of variance that contribute to the differences we have observed.

There are several contributions in this paper. In Table 1, we provide a useful summary of several published empirical studies on document relevance judging by multiple judges. We believe we are the first to argue for using Cohen’s κ , which corrects for agreement by chance, to measure inter-judge agreement in IR relevance assessment. We demonstrate empirically using Kendall’s τ and RMS error that inferred measures (for non-uniform random sampled judging) correlate well with AP and NDCG both by topic and averaged over all topics. On the data available, topic-based sampling appears a better approach than effort-based sampling. Finally, we have tried to control for degrees of task and topic expertise in our groups of judges to investigate their effect.

6. CONCLUSIONS

The Cranfield (and TREC) methodology gives us a user model. In an experiment that is intended to model a real-world search scenario, such as the TREC Enterprise track, it is desirable that judgements in the model correspond to opinions of a real user. Past studies have shown that different judges from the same population are exchangeable. Our judges are drawn from different populations.

Our conclusions arising from the experiments conducted are as follows. First, we have investigated whether task and/or topical expertise affect relevance judgements in a consistent manner — they do. Relatively low levels of agreement exist between types of judge, and the agreement is even less between gold and bronze standard than gold and silver.

Second, we have investigated whether such differences in relevance assessments affect performance scores — again, they do. On a per topic basis, these differences definitely affect the scores on the inferred measures. Per topic variation is well known to occur in the Cranfield method, leading to the reporting of averaged measures over a sufficient numbers of topics. In our investigation, when averaged over the 33 topics that were judged by all three categories of judges, these differences affect the measures in a small but consistent way. Bronze standard judges appear to be less discerning than gold standard judges leading to higher overall scores. Conversely, system scores from the silver standard judge's judgements were slightly worse than those from the gold standard judges.

Overall we conclude, like earlier investigators, that the Cranfield method of evaluation is somewhat robust to variations in relevance judgements. Having controlled for task and topic expertise, system performance measures show statistically significant, but not large, differences. Similarly, system orderings allow us to identify “good” and “bad” IR systems at a broad-brush level.

However, we find that bronze standard judges may not be a reliable substitute for the gold standard task and topic experts. Neither silver standard nor bronze standard judges were the topic originators, yet silver standard judges performed closer to gold standard than did bronze standard judges. This suggests that disagreement stems not just from judging by non-originators. It is possible that unfamiliarity with task and topic context plays a major role.

When evaluating relative system performance for task-specific IR, it can be important to obtain judgements from either the gold standard judges or a close surrogate to accurately reflect user preferences. The new sampling methods and inferred measures discussed here allow this to be done at lower cost and effort than before.

7. ACKNOWLEDGMENTS

We would like to thank the CSIRO science communicators, TREC Enterprise 2007 participants, Simon Barry, Lish Fejer, Donna Harman, David Hawking, Stephen Robertson, Wouter Weerkamp, and other colleagues and reviewers who provided advice and assistance.

8. REFERENCES

- [1] R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, to appear.
- [2] J. A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In *Proc. SIGIR*, 2006.
- [3] P. Bailey, N. Craswell, I. Soboroff, and A. P. de Vries. The CSIRO enterprise search test collection. *SIGIR Forum*, 41(2), December 2007.
- [4] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proc. SIGIR*, 2004.
- [5] R. Burgin. Variations in relevance judgments and the evaluation of retrieval performance. *Information Processing & Management*, 28(5):619–627, Sep-Oct 1992.
- [6] J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.
- [7] C. W. Cleverdon. The effect of variations in relevance assessments in comparative experimental tests of index languages. Technical Report ASLIB part 2, Cranfield Institute of Technology, 1970.
- [8] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.
- [9] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient construction of large test collections. In *Proc. SIGIR*, 1998.
- [10] B. D. Eugenio and M. Glass. The kappa statistic: a second look. *Computational Linguistics*, 30(1):95–101, 2004.
- [11] S. P. Harter. Variations in relevance assessments and the measurement of retrieval effectiveness. *JASIS*, 47(1):37–49, 1996.
- [12] K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *Proc. SIGIR*, 2000.
- [13] K. S. Jones and K. van Rijsbergen. Information retrieval test collections. *Journal of Documentation*, 32:59–75, 1976.
- [14] M. E. Lesk and G. Salton. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*, 4:343–359, 1969.
- [15] S. Mizzaro. Measuring the agreement among relevance judges. In *Proc. MIRA 99: Evaluating Interactive Information Retrieval*, April 1999.
- [16] R. Rietveld and R. van Hout. *Statistical Techniques for the Study of Language and Language Behaviour*. Mouton de Gruyter, 1993.
- [17] S. Sigel and N. J. Castellan. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, 1988.
- [18] I. Soboroff. A comparison of pooled and sampled relevance judgments. In *Proc. SIGIR*, 2007.
- [19] E. Sormunen. Liberal relevance criteria of TREC: counting on negligible documents? In *Proc. SIGIR*, 2002.
- [20] A. Trotman and D. Jenkinson. IR Evaluation Using Multiple Assessors per Topic. In *Proc. ADCS*, 2007.
- [21] A. Trotman, N. Pharo, and D. Jenkinson. Can we at least agree on something? In *Proc. SIGIR Workshop on Focused Retrieval*, 2007.
- [22] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proc. SIGIR*, 1998.
- [23] E. M. Voorhees and D. Harman. Overview of the Fifth Text REtrieval Conference (TREC-5). NIST, 1996.
- [24] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proc. CIKM*, 2006.
- [25] E. Yilmaz, E. Kanoulas, and J. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In *Proc. SIGIR*, 2008.