Including Summaries in System Evaluation

Andrew Turpin† Falk Scholer† Kalvero Järvelin* Mingfang Wu† J. Shane Culpepper†

†RMIT University, Melbourne, Australia
{andrew.turpin,falk.scholer,mingfang.wu
shane.culpepper}@rmit.edu.au

*Uni. of Tampere, Tampere, Finland
kalervo.jarvelin@uta.fi

ABSTRACT

In batch evaluation of retrieval systems, performance is calculated based on predetermined relevance judgements applied to a list of documents returned by the system for a query. This evaluation paradigm, however, ignores the current standard operation of search systems which require the user to view summaries of documents prior to reading the documents themselves.

In this paper we modify the popular IR metrics MAP and P@10 to incorporate the summary reading step of the search process, and study the effects on system rankings using TREC data. Based on a user study, we establish likely disagreements between relevance judgements of summaries and of documents, and use these values to seed simulations of summary relevance in the TREC data. Re-evaluating the runs submitted to the TREC Web Track, we find the average correlation between system rankings and the original TREC rankings is 0.8 (Kendall τ), which is lower than commonly accepted for system orderings to be considered equivalent. The system that has the highest MAP in TREC generally remains amongst the highest MAP systems when summaries are taken into account, but many other systems become equivalent to the top ranked system depending on the simulated summary relevance.

Given that system orderings alter when summaries are taken into account, the small amount of effort required to judge summaries in addition to documents (19 seconds vs 88 seconds on average in our data) should be undertaken when constructing test collections.

Categories and Subject Descriptors

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

General Terms

Experimentation, Measurement, Performance

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.

1. INTRODUCTION

The Cranfield methodology, established in the 1960s [6], is still the most popular way to evaluate Information Retrieval systems, particularly given the existence of the TREC collections [27]. The success of this experimental paradigm, which is used much more widely than user-based experiments, is often attributed to a number of factors: results are easily reproduced, and so can be replicated by other research groups and used for comparative studies; it is typically much less time consuming and costly than user experimentation; and, it has a strong history, having been the dominant paradigm of system evaluation for over 30 years.

The Cranfield methodology uses a relevance score for each document-topic pair to reduce a list of documents returned by an IR system to an ordered vector of relevance levels. In turn, this list of relevance levels is summarized using a metric such as Average Precision [4], *bpref* [3], or nDCG [12]. These metrics are then averaged over a set of topics to give a single score for a system, and systems can then be ranked, relative to each other, by these mean metric values.

Recent research has shown that, even if System A has statistically significantly higher mean metric values than System B, there is no guarantee that users will perform their tasks better with System A than System B [1, 2, 10, 22, 23]. Some studies show that users prefer System A to System B, even though their performance is not enhanced [1, 14].

One possible reason for the differences in system ranking using batch and user experiments is that current IR systems actually require users to make two decisions: one on a short summary, and another on the document itself. That is, users are typically presented with a list of summaries of retrieved documents, and only proceed to examine a particular document itself if they find the summary appealing. The triage process of examining the document summary is ignored in the Cranfield methodology. There is an implicit assumption that all summaries presented to the user will accurately reflect the underlying document, or that all documents are always examined, whatever the summary. Given this assumption, evaluating systems using lists of document relevance levels is valid. As we have all experienced, however, the document summary presented by a search system often does not accurately reflect the document's content relative to the information need posed.

In this paper we explicitly examine the effect of including the summary examination stage of the retrieval process in batch evaluations. We first present data from a small user study examining the rate with which summaries do not reflect the underlying document for a topic. These results

support the intuition that often summaries do not lead users to correct choices relative to document relevance. Given that users may miss some relevant documents in a ranked list, even though these documents may have helped to fulfill their information need, we re-evaluate TREC Web Track runs taking this into account. As TREC data does not include judgements about whether a document might be selected based on its summary, we simulate such judgements using parameters from our user study.

Our results indicate that the ranking of systems differ from that found in the TREC runs (average Kendall's $\tau = 0.8$) by more than the level that is commonly accepted to show that two system rankings are equivalent ($\tau = 0.9$ [25]). Moreover, while the system that has the highest mean average precision (MAP) score in the original TREC runs generally remains within the top 10 ranks, more systems become statistically equivalent to the top run than when summaries are not included.

2. BACKGROUND

In this paper, we investigate the effect that the incorporation of a summary reading step – a common step when conducting a search with most text-based information retrieval systems – has on the evaluation process. To our knowledge, no one has explicitly studied the effect on system orderings in a batch experiment if the summary reading stage of retrieval is taken into account. While the creation of document summaries from various sources is a research field in its own right (in particular, there have been a few papers describing techniques for generating query-biased summaries such as those commonly used for web retrieval, see for example Tombros and Sanderson [21], Scholer and Williams [18], Varlamis and Stamou [24], and Wu et al. [28]) these papers do not concern themselves with batch system evaluation, but instead aim to analyse the effectiveness of different summary creation approaches. This paper is not an investigation of summary generation efficiency or effectiveness.

In the batch-style evaluation of IR systems, the performance of individual retrieval systems can be calculated using a metric such as MAP, based on a fixed set of topic-document relevance judgements [27]. This metric gives an overall ordering of relative system performance. Kendall's τ is a measure of correlation, and shows the strength of the relationship between two rankings based on the number of pairwise swaps that is required to transform one ranking into another [19]. τ was used to quantify the level of agreement between system rankings by Voorhees [25], who investigated the effect that the use of different relevance judgements has on the outcome of batch experiments. In a series of experiments, the relevance of documents in the TREC-4 *ad hoc* collection was re-judged by multiple TREC assessors, while documents from the TREC-6 *ad hoc* collection were re-judged by students from the University of Waterloo. The results showed that the correlation between system rankings based on judgements by different assessors is generally around a level of $\tau = 0.9$ or greater. Hence, different system rankings with a correlation greater than 0.9 have been considered equivalent in subsequent papers (for example, Carterette and Allan [5], Sanderson et al. [16], Voorhees [26], and Yilmaz and Aslam [29]). Although some work has suggested that applying absolute thresholds to correlations may be problematic [17], the use of Kendall's τ is the current standard for comparing the difference between system orderings in retrieval evaluation.

Many batch experiments, and many of the system performance metrics used in this evaluation approach, consider relevance on a binary scale: a document is either relevant, or it is not. However, investigations into the use of multiple-level relevance in the TREC framework have suggested that considering different levels of relevance can offer additional insight into the way in which users might interact with documents [12]. As a consequence, this can also affect their view of system performance. For example, Sormunen analysed documents returned in response to topics from the TREC-7 and TREC-8 adhoc task, and classified into a four-level relevance scale. Results indicated that nearly half of the documents that are judged "relevant" by TREC judges are only in fact *marginally* relevant, containing no information beyond what is already specified in the user's information need [20]. We make use of a four-point relevance scale in our user study on the impact of a summarization step on retrieval evaluation. Explicitly separating highly relevant documents will allow us to examine whether the summaries produced for such documents are judged relevant by users more often than those produced for documents that are only judged marginally relevant.

3. FRAMEWORK

For a given query and search engine, let R be the list of relevance levels of the documents, in ranked order, returned by the engine for that query. The ordered list R, with position index i, can be reduced to a single number for each query-engine pair in a variety of ways. For batch experiment evaluation, the elements in R are often mapped to binary values:

$$R'_i = \begin{cases} 0, & R_i = 0 \\ 1, & \text{otherwise,} \end{cases} \quad (1)$$

and then summarized. For example, Precision At n is simply the proportion of non-zero entries in the first n elements of R',

$$\text{P@}n = 1/n \times \sum_{i=1}^{n} R'_i \quad (2)$$

Average precision for a list of results is the sum of P@i, where i is the index of all non-zero entries in R', normalized by $\mathcal{R}$, the total number of relevant documents that exist for the query:

$$\text{AP} = \frac{1}{|\mathcal{R}|} \sum_{i=1}^{|R'|} R'_i \times \text{P@}i \quad (3)$$

The mean of these values over many queries is usually reported, with Mean Average Precision (MAP) being the most commonly used metric in TREC.

To allow for the extra level of decision made by users when presented with short summaries for documents in a results list, we define an ordered list C, where $C_i = 1$ if users would select the summary for document i in order to read the full document, and $C_i = 0$ if users would not choose to read the full document.

1	cause of petrol price change
2	kangaroo population control
3	van gogh work
4	charlotte bronte jane eyre
5	space shuttle program scandal
6	is there an australian recession
7	cost green internet
8	bigfoot scientific proof
9	500 card bidding strategy
10	world of warcraft addiction evidence
11	health blueberries
12	nokia factory closing germany

Table 1: Queries used in this study. Each query was accompanied with an expanded description of the information need underlying that query.

When the summary step is taken into account, a relevant document ($R_i > 0$) should only make a positive contribution to a system performance metric if the user would actually view that document ($C_i = 1$). Hence we redefine

$$R'_i = \begin{cases} 0, & R_i = 0 \text{ or } C_i = 0 \\ 1, & R_i > 0 \text{ and } C_i > 0. \end{cases} \quad (4)$$

If we use the formulation of R' in Equation 4 for calculating P@1, for example, it can be seen that a system that returns a relevant document (according to the judgements used in the evaluation) at rank 1 that also has a good summary (that leads the user to view the document) will score P@1=1, while a system that returns a relevant document at rank 1 that has a bad summary will score P@1=0. Without taking the summaries into account, both systems would score P@1=1. Similarly, when comparing systems with other metrics, it is possible that incorporating C_i into the metric will change its value. More importantly from the perspective of system evaluations, we are interested in whether the introduction of C_i will change the relative ordering of system performance.

Note that this framework assumes that we are evaluating systems that are useful for fulfilling information needs that cannot be satisfied with simple factoids that could be found in short document summaries. Our assumption is that the systems are to be used for answering information needs that are more complex than navigational queries, or simple question-answering style queries, hence require investigation of the documents themselves to be satisfied.

4. ARE RESULTS LIKELY TO CHANGE?

To investigate whether it is likely that including C_i in metrics will alter batch evaluation results, we conducted a small user study, with the five authors of this paper as participants. Twelve queries were suggested by the participants, listed in Table 1. In order to get documents and their corresponding summaries, we submitted each query to Google, assuming their summary generation and document ranking algorithms were likely to represent a good information retrieval system. Fifteen documents, together with their corresponding summaries, were obtained per topic. The "owner" of each query removed duplicate documents, images, videos and Wikipedia entries from Google's lists, and also ensured that there were likely to be a good spread of document relevance levels (in their opinion). That is, queries that resulted in lists with only highly relevant documents, or conversely almost none, were not chosen for this study. The owner also wrote a description of an information need to accompany the query, similar in style to a TREC narrative.

The owner and three other participants (hereafter referred to as "users") judged the summaries for the resulting 15 documents for each topic on a binary scale, answering the question "would you click on the link to view the underlying document to answer this information need". Users were asked to judge all 15 summaries, in order, and to judge them independently of one another. To additionally account for possible learning effects, the original summary list was divided into three *blocks*, which were rotated so that each user saw the list in a different order.

The documents themselves were again judged by the owner and three users, based on the following 4-point relevance scale.

0: Irrelevant: the document does not contain any information about the topic.

1: Marginally Relevant: the document mentions in passing the theme or any aspect of the topic, or provides only links to highly relevant or fairly relevant documents.

2: Fairly relevant: The document discusses the topic or some aspect(s) of the topic, but the discussion is not exhaustive.

3: Highly Relevant: the document discusses the topic exhaustively. In case of a multiple aspect topic, the document discusses either all or any aspect of the topic exhaustively.

To reduce possible confounds due to ordering effects, all documents were shown to users in the same order in which they were presented with summaries (that is, different users saw lists for topics in different orders). In addition, to allow for the possibility of learning between the summary and document judging phases, a one week gap was enforced between each phase.

As explained above, both summaries and documents were judged by the owner (topic creator) and three users. From these observations, we derive two sets of relevance judgements: *owner* judgements are those made by the topic creator. This is roughly analogous to how relevance judgements are made in the standard TREC framework: the topic creator is also responsible for determining the relevance of candidate answer documents returned by a search system. A second set of relevance judgements, *user*, is derived by taking the majority judgement assigned by the three user judges. In the few cases where there was no majority, the mean judgement rating was used, rounded to the nearest whole number.

The raw agreement between the owner and user judgement sets for summary relevance is 73.9%. For document relevance, raw agreement is 59.4%; this level seems comparatively low, but documents are rated on a four-point relevance scale. When the document judgements are collapsed to a binary scale using Equation 1, then agreement on document judgements is 74%.

Figure 1 shows the proportion of documents in each of the four relevance categories (majority user) whose summaries

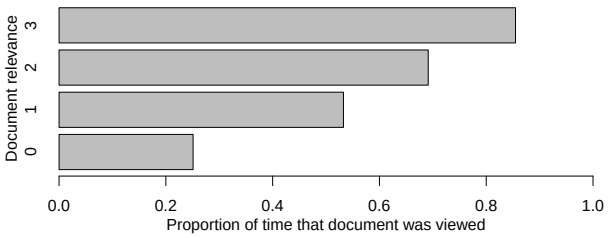

Figure 1: The proportion of documents in each relevance category that received a $C_i = 1$ judgement over all judgements made by users.

	None	User C	Owner C
Owner R	0.66	0.40 (-39%)	0.31 (-53%)
User R	0.52	0.29 (-44%)	0.26 (-50%)

Table 2: MAP over the 12 topics using different definitions of R and C. MAP is calculated using: all $C_i = 1$ (None); majority of user C judgements (User); and the owner's C (Owner). Percentages indicate reduction from the MAP in column None.

were judged "clickable" by the users. For example, when presented with a summary for a highly relevant document, users decide that they would proceed to view the underlying document 86% of the time. As can be seen, there is a strong positive relationship between the quality of the underlying document, and the likelihood with which the document's summary will actually lead a user to view the underlying resource (that is, there is a positive correlation between R_i and C_i). This is reassuring, since it indicates that summaries are, in general, doing their job!

Based on our framework, we investigate the effect of incorporating the summary step on system performance metrics. Table 2 shows the results of computing MAP over the 12 topics on the lists judged by users using Equation 3 and the revised definition of R' given in Equation 4. Rows show two different types of document relevance (R) based on the judgements of the topic developer (*owner*), or the majority of users (*user*). Columns indicate the source of summary judgement information: none, indicating the full MAP score with summaries not taken into account; and summary judgements as determined by the user and owner groups. In all cases, the introduction of the summary step changes MAP scores (a fall of at least 39%).

Absolute MAP scores are not very enlightening as indicators of system performance [27]. However, relative performance scores should be interpretable in a meaningful way. For example, running a single system over a set of 12 topics would give a different performance metric (such as average precision) for each topic. The relative ease with which the system can find answers for each topic is thus determined.

To study the robustness of batch evaluation of these topics, we investigate whether the same relative topic difficulties are retained after a summary judging step is incorporated into the evaluations. If the conclusions from the batch framework are robust, we would expect the orderings to be largely similar.

Table 3 shows the Kendall's τ correlation between topic

Ordering 1		Ordering 2		τ	Description
C	R	C	R		
none	user	none	owner	0.73	Just docs
none	user	user	user	0.52	User judges
none	owner	owner	owner	0.67	Owner judges
none	owner	user	user	0.61	Interactive
none	owner	user	owner	0.61	Ad hoc

Table 3: Correlations (Kendall's τ) between two orderings of the 12 topics using the MAP metric, where each ordering uses a different combination of C and R. The description column is explained in the text.

orderings, based on different sources of document and summary relevance judgements. The configuration for Ordering 1 is shown in the first two columns: Ordering 1 does not make use of any summary judgements (that is, the difficulty ordering is based only on document relevance judgements, sourced from the owner, or the majority of users). This is correlated against Ordering 2 described in the next two columns.

The first row shows the correlation between topic orderings that arises only from a different source of document relevance (user and owner), with no consideration of the summary step, leading to a τ of 0.73. It is understandable that the MAP will change for topics when different judges are employed because judges agree anywhere from around 60% [7] to 74% as in this study. This τ value, therefore, can be interpreted as the level of background noise in the orderings, arising from disagreements between judges. The second row demonstrates the effect of incorporating a summary step, when document judgements are held constant (based on users). The introduction of the summary step leads to substantially higher perturbation in the topic orderings ($\tau = 0.52$) than was observed by changing the source of document relevance judgements. The effect of introducing the summary step when using owner judgements is shown in the third row; again the correlation is substantially lower than when changing document relevance (although not as low as for user judgements). The fourth row indicates a situation that is similar to that encountered in the TREC Interactive Tracks: the system performance based only on the document judgements of the owner (topic creator) is compared with the user's own perception of document relevance, tempered by their behaviour when faced with summaries in the initial search results list. The correlation of topic orderings is again below that seen when changing document relevance judgements alone. The final row of the table is similar to what might be encountered if a TREC ad hoc-style task was judged using an actual search system that returns summaries: the pure document relevance as decided by the owner is convolved with the user's behaviour when first presented with a summary of the underlying document.

As can be seen from this analysis of our user study data, the introduction of a summary step in the batch evaluation process leads to substantial falls in the correlation of topic difficulty orderings, beyond the level that is observed when simply swapping document relevance judgements. We now examine how accounting for summaries might impact on relative system orderings.

1. For each automatic run in the set of runs (T9, T10) do
   For each document d in the run do
   Set $r \leftarrow$ relevance of d in qrel file.
   Set $t \leftarrow$ random number between 0 and 1.
   If $r > 0$ and $t > p(r)$ then
   Replace d with some non-relevant document.
2. Calculate metrics on the modified runs.
3. Calculate Kendall's tau between the original system ordering in the runs and the modified runs.

Figure 2: The procedure used to generate and evaluate TREC runs taking C_i into account.

5. RE-EVALUATING TREC-9 & TREC-10

A good search engine should produce summaries that will invite users to select highly relevant documents, and to avoid irrelevant documents. As such, there should be a strong correlation between R and C. From the data presented in the previous section we can see that this is the case for Google on the 12 topics we studied. If we are to re-evaluate system ordering in TREC, we need some way of predicting C_i from R_i, as summaries are not considered at all within TREC ad hoc tasks.

A simple first approach is to assume that the likelihood of selecting a document is determined by a simple Bernoulli trial (coin toss), with the mean probability of selecting determined by the underlying document's relevance:

$$\hat{C}_i \stackrel{d}{=} \text{Bi}(p(R_i)), \tag{5}$$

where $p(x)$ gives the probability of selecting a summary extracted from a document of relevance level x. For example, if we use the empirical data gathered in Section 4 of this paper (and assume that our relevance category 2 and 3 maps to TREC's category 2),

$$p(R') = \begin{cases} 0.25, & R' = 0 \\ 0.53, & R' = 1 \\ 0.77, & R' = 2. \end{cases} \tag{6}$$

Note that we are adopting the three TREC relevance levels for R' as we are particularly interested in evaluating TREC data, and use the standard TREC assumption that TREC category 1 documents are relevant when computing metrics.

Using Equations 5 and 6, we can assume some C values for any given set of TREC documents, and apply Equation 4 to recompute the metrics for a particular *TREC run*. A TREC run is a list of documents returned for a given topic by some system. Then, for a set of runs, we can order the systems based on the new metric values and see how they compare to the original order of the runs using Kendall's τ to measure the correlation between the two. Because there is randomness involved in determining C_i from R_i using this method, we repeat the process 1000 times, and look at the distribution of τ values. The algorithm in Figure 2 describes the process for generating one system ordering from TREC runs.

Figure 3 shows the Kendall's τ values between the system ordering given by the MAP values of the original TREC-9 and TREC-10 runs and each of the 1000 generated orderings. In each case all but the manual runs were included, giving 40 runs for TREC-9 and 77 runs for TREC-10. Each run represents up to 1000 documents per topic for 50 topics.

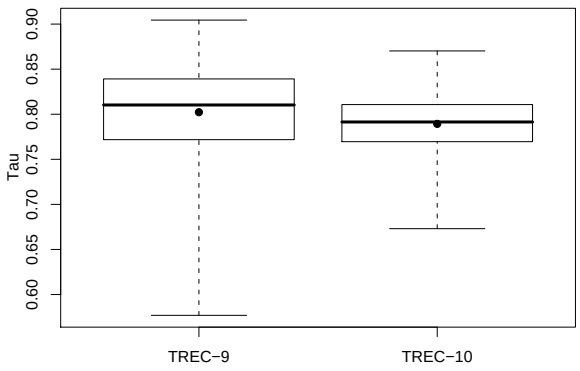

Figure 3: Correlation between the system ordering based on MAP given by TREC and each of the system orderings given by 1000 simulated runs (Kendall's τ).

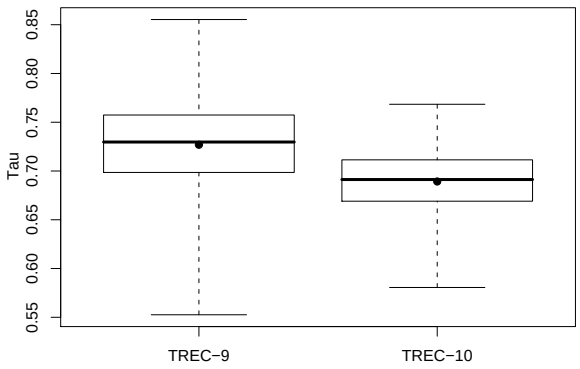

Figure 4: Correlation between the system ordering based on P@10 given by TREC and each of the system orderings given by 1000 simulated runs (Kendall's τ).

The boxes indicate the 25th and 75th percentiles of the values, the solid line the median, the whiskers show the extreme values, and the mean is shown as a dot. As can be seen, nearly all of the system orderings generated when taking summary relevance into account are less than $\tau = 0.9$, the level that can be expected when using different documents relevance judgements to evaluate runs [26]. The 95th percentile for the τ values is 0.87 for TREC-9, and 0.84 for TREC-10, both less than the 0.9 level.

We also analyse the τ values for P@10, shown in Figure 4. The correlations of system orderings are lower for both TREC-9 and 10 than they were for MAP. Again, the values are substantially lower than the threshold of 0.9, but this threshold was established using the MAP metric so should be treated with caution when applied to P@10 data.

While Figures 3 and 4 show that overall system rankings alter, perhaps of more interest is what happens to the top ranked system in the TREC runs: does it remain at the top in the simulated orderings? Figure 5 shows that the top ranked TREC system based on MAP remains in the top 10 systems for nearly all simulated runs on both collections. Given that often there is no statistically significant difference

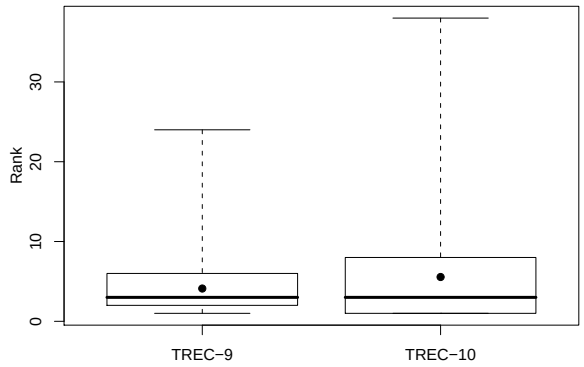

Figure 5: Position of the system with the highest MAP in TREC in each of the 1000 system orderings of the simulations.

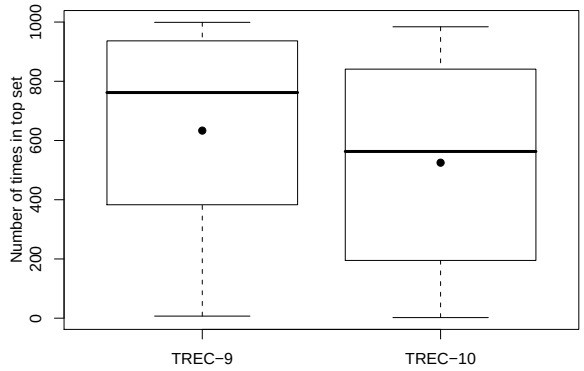

Figure 6: Number of times the MAP of each system is equivalent (t-test, $p < 0.05$) to the system with the highest MAP in the 1000 simulated runs.

between several of the top ranked systems, we also computed the number of times that the TREC top ranked system was statistically significantly worse than the top ranked system in any of the simulations. We define the *top set* of runs as the systems that have metric values that are not significantly different than the highest ranked system according to a t-test using $p = 0.05$. The top TREC-9 system [8] was not in the top set of the 1000 simulated runs 13 times, and the top TREC-10 system [9] was not in the top set 16 times.

While the system with the highest MAP value in a TREC run generally remains in the top set of the simulated runs, different systems enter the top set in different simulations. Figure 6 shows the distribution of the number of times a system is in the top set. The median value of 762 for TREC-9 indicates that half of the systems are in the top set in 762/1000=76.2% of the simulations; based on the mean, a system is in the top set 633/1000=63.3% of the time. For TREC-10, the median and mean values are 563 and 525, respectively.

One possible reason for the expansion of the top set is that MAP values decrease when summaries are taken into account, and so the range of possible MAP values shrinks. This, in turn, might remove statistically significant differences between the highest MAP value and others. Figure 7 shows the distribution of the number of systems that are in

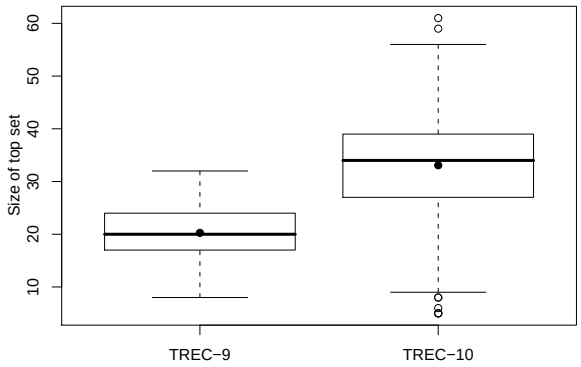

Figure 7: Number of systems that are statistically equivalent to the highest MAP in the 1000 simulated runs: the size of the top set.

the top set for the 1000 simulated runs. The original TREC-9 runs has a top set size of 15, and from the figure it is closer to 20, on average, when summaries are included. Likewise, for TREC-10, the original top set size is 28, while it is 35 on average when summaries are included.

6. DISCUSSION

The user data in this paper shows that 14% of highly relevant and 31% of relevant documents are never examined because there summary is judged irrelevant (Figure 1). Given that most modern search engines display some sort of summary to the user, it seems unrealistic to judge system performance based on experiments using document relevance judgements alone. Our re-evaluation of TREC data confirms that systems rankings alter when summary relevance judgements are added (Figure 3).

The re-evaluation is based on a simple probabilistic model of summary relevance using document relevance and the proportions in Equation 6 from our user study. Given that the user study has only 5 participants and 12 topics, these proportions should not be heavily relied upon, and are merely indicative of the type of values one might expect. In order to test how robust the re-evaluation is to the choice of proportions, we ran additional simulations, following the algorithm in Figure 2, and setting the probabilities of reading a document after viewing its summary to a range of values. More formally, we tested all p functions in the family

$$p(R') = \begin{cases} \alpha, & R' = 1 \\ \beta, & R' = 2 \end{cases}$$

for all combinations of $\alpha, \beta \in \{0.45, 0.50, \ldots, 1.00\}$. A contour plot of mean τ values (10 simulations per location), showing the correlation between system orderings for a simulation with a particular combination of parameters and the original TREC-9 system ordering, is shown in Figure 8. The contours are interpolated between the discrete data points [15]. For example, when the probability for viewing a $R' = 2$ document is 0.8, and the probability of viewing a $R' = 1$ document is 0.6, then the correlation between this system ranking and the original data that does not take the summary step into account is around $\tau = 0.85$.

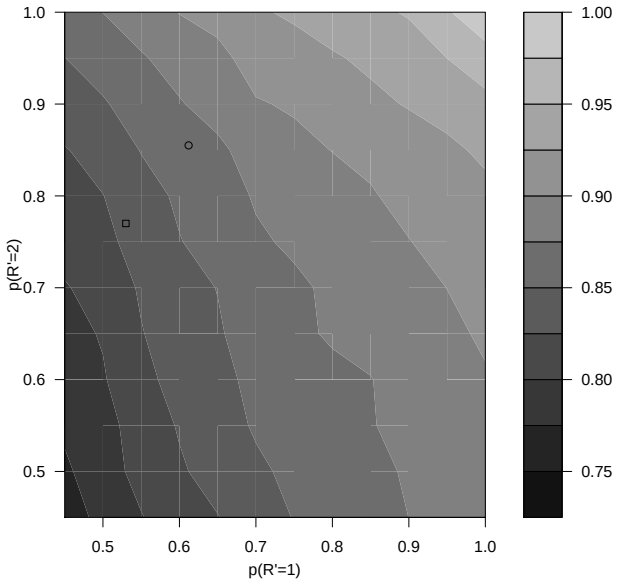

Figure 8: Contour plot of mean τ values for 10 simulated runs ranked with MAP for various choices for $p(R')$. The intersection of grid lines mark actual data points, with the contour areas interpolated. The square and circle indicate the combination of parameters derived from our user study.

The square on the plot indicates the parameter values used in this study. When mapping the four proportions from our user data shown in Figure 1 to the three TREC relevance categories, we chose to average category 2 and 3, to get $p(2) = 0.77$ as shown in Equation 6. If we choose to fold our four point relevance scale slightly differently by averaging category 1 and 2, to get $p(1)$, and set $p(2)$ to be the proportion for category3, then when we reevaluate TREC-9 runs we get the mean τ shown by the circle in Figure 8.

It seems, therefore, for the summary step to not have an impact on system ordering ($\tau > 0.9$—the top right corner of the contour plot), the summaries would need to be extremely accurate in how they reflect the content of the underlying document, for a given information need.

Varying $p(1)$ has much more effect on system ordering than varying $p(2)$ because TREC judgements generally contain many more relevant documents than highly relevant documents. It is not obvious what would happen if TREC runs were scored assuming that category 1 documents were irrelevant. There would be many less relevant documents, so metric values would change, and a single change in relevance due to summary usage may have a large affect on metric values. It is also unclear how user judgements of summaries might change if their task was to find highly relevant documents.

While our approach to assigning C to existing TREC data is simple, but robust, it is far from perfect. Ideally, every system submitted to TREC should produce both a summary and a document so that both can be judged by NIST assessors. Judging summaries is much faster than judging documents: on average our participants took 19 seconds to judge a summary, but 88 seconds to judge a document. A less ideal, but perhaps more practical approach, would be for a single summary to be produced for any document-topic pair by a reference system, and that summary judged by the collection creators.

One of the main criticisms of the Cranfield methodology, and papers based on TREC data, is that the reliance on relevance judgements computed by NIST assessors does not allow ready translation of the results into the real world where judgements are made by different people [2, 11, 22, 23]. In this paper we have apparently compounded this problem by injecting a second relevance judgement per document into collection construction: binary judgements of short summaries. Relying on both R_i and C_i judgements in a collection for evaluating systems may increase the chance that batch and user experiments will not concur on system *ordering*. We argue that without including the C_i factor, there is even less chance that batch and user experiments will agree on system *evaluation*. Users of current search engines are making decisions on both summaries and documents, and so both must be included in batch evaluations if they are to attempt to model reality. It is better to include the C_i factor and attempt to solve the new problem of matching it to actual user behaviour, than to not include it at all and remain with old problems of batch and user experiment mismatch.

One side effect of adding an extra judgement step into the batch evaluation paradigm may be that system ranking algorithms will be altered to not only return documents that appear to match the information need of the query, but to also prefer documents that can generate good summaries (that is, documents that can be summarized accurately for that particular information need). We believe that this is a potential area for improvement in current search systems, and are investigating such algorithms in our lab.

In the user study for this paper, we used summaries generated from Google, which we presume are close to the current state-of-the-art. Hence our estimate of the "noise" introduced into the batch evaluation by summaries may be conservative. As summary generation algorithms improve, the effect of summaries on document viewing behavior will change. However, if summaries are judged explicitly as part of the batch evaluation paradigm, this kind of variation will be accurately included in system rankings.

The simulated summary relevance values we used to re-evaluate TREC runs are based on document relevance levels. There are other factors that might influence the way users react to particular summaries. A "trust bias" effect has been demonstrated in results returned by web search engines [13], where users trust the search system to return useful items early in the ranking, so are more likely to select documents that appear near the top of the list. There may also be an "authority" effect; for example, a user looking for economic indicators about a recession might choose to avoid viewing a result from `www.my-ranting-blog.net`, no matter how pertinent the summary text looks. While our simulations do not explicitly include these factors, the results in Table 2 show that MAP decreases significantly – and with enough variability to cause re-orderings of relative per-topic performance – in the user data, where these factors are implicitly included. Therefore, our approximation of summary-selecting behavior based on relevance levels should reflect such biases to some extent. We plan to investigate more complex models of user behaviour when interacting with summaries in future work.

7. CONCLUSIONS

We performed a small-scale user study of a two step relevance assessment process of web documents for non-factoid information needs, with both summaries and documents judged. The results indicated that the summary evaluation step clearly influences which documents are seen as relevant. This seems to depend on the degree of document relevance: highly relevant documents yield summaries that users are more likely to perceive as relevant than fairly or marginally relevant documents (86%, 69%, and 53% respectively). Using these probabilities for simulation, we re-evaluated TREC-9 and TREC-10 Web Track runs and examined the effect of summaries on system ordering. We observed that the system that has the highest MAP in TREC generally remains amongst the highest MAP systems when summaries are taken into account, but many other systems become equivalent to the top ranked system depending on the simulated summary relevance. For the summary evaluation step not to have a significant impact on system ordering, document summaries would need to correctly lead people to select a relevant document for reading at least 80% of the time. Therefore, to foster IR system development toward producing better summaries, and IR evaluation toward more realism, test collections for evaluating information retrieval systems should include both summary and document evaluation judgements.

8. REFERENCES

[1] A. Al-Maskari, M. Sanderson, P. Clough, and E. Airio. The good and the bad system: does the test collection predict users' effectiveness? In *Proc. ACM SIGIR*, pages 59–66, Singapore, Singapore, 2008.

[2] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? In *Proc. ACM SIGIR*, pages 433–440, Salvador, Brazil, 2005.

[3] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proc. ACM SIGIR*, pages 25–32, Sheffield, UK, 2004.

[4] C. Buckley and E. M. Voorhees. Retrieval system evaluation. In Ellen M. Voorhees and Donna K. Harman, editors, *TREC: experiment and evaluation in information retrieval*. MIT Press, 2005.

[5] B. Carterette and J. Allan. Incremental test collections. In *Proc. ACM CIKM*, pages 680–687, Bremen, Germany, 2005.

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

[7] C. Cleverdon. Optimizing convenient online access to bibliographic databases. *Information Services and Use*, 4(1-2):37–47, 1984.

[8] D. Hawking. Overview of the TREC-9 Web track. In *TREC-9*, pages 87–102, Gaithersburg, MD, 2000.

[9] D. Hawking and N. Craswell. Overview of TREC 2001 Web track. In *TREC 2001*, pages 61–67, Gaithersburg, MD, 2001.

[10] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kraemer, L. Sacherek, and D. Olson. Do batch and user evaluations give the same results? In *Proc. ACM SIGIR*, pages 17–24, Athens, Greece, 2000.

[11] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*. Kluwer Academic Publishers, 2005.

[12] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[13] T. Joachims, L. A. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proc. ACM SIGIR*, pages 154–161, Salvador, Brazil, 2005.

[14] D. Kelly, X. Fu, and C. Shah. Effects of rank and precision of search results on users' evaluations of system performance. Technical Report TR-2007-02, University of North Carolina, 2007.

[15] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[16] M. Sanderson and H. Joho. Forming test collections with no system pooling. In *Proc. ACM SIGIR*, pages 33–40, Sheffield, United Kingdom, 2004.

[17] M. Sanderson and I. Soboroff. Problems with Kendall's tau. In *Proc. ACM SIGIR*, pages 839–840, Amsterdam, The Netherlands, 2007.

[18] F. Scholer and H. E. Williams. Query association for effective retrieval. In *Proc. ACM CIKM*, pages 324–331, McLean, VA, 2002.

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

[20] E. Sormunen. Liberal relevance criteria of TREC – counting on negligible documents? In *Proc. ACM SIGIR*, pages 324–330, Tampere, Finland, 2002.

[21] A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In *Proc. ACM SIGIR*, pages 2–10, Melbourne, Australia, 1998.

[22] A. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *Proc. ACM SIGIR*, pages 225–231, New Orleans, LA, 2001.

[23] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *Proc. ACM SIGIR*, pages 11–18, Seattle, WA, 2006.

[24] I. Varlamis and S. Stamou. Semantically driven snippet selection for supporting focused web searches. *Data & Knowledge Engineering*, 68:261–277, 2009.

[25] E. M. Voorhees. Variations in relevance judgements and the measurement of retrieval effectiveness. In *Proc. ACM SIGIR*, pages 315–323, Melbourne, Australia, 1998.

[26] E. M. Voorhees. Evaluation by highly relevant documents. In *Proc. ACM SIGIR*, pages 74–82, New Orleans, LA, 2001.

[27] E. M. Voorhees and D. K. Harman. *TREC : experiment and evaluation in information retrieval*. MIT Press, 2005.

[28] M. Wu, M. Fuller, and R. Wilkinson. Searcher performance in question answering. In Proc. ACM SIGIR, pages 375–381, New Orleans, LA, 2001.

[29] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In Proc. ACM CIKM, pages 102–111, Arlington, VA, 2006.