Modelling Disagreement Between Judges for Information Retrieval System Evaluation

Andrew Turpin *Falk Scholer*

School of Computer Science & IT
RMIT University
GPO Box 2476
Melbourne 3001

{*andrew.turpin,falk.scholer*}*@rmit.edu.au*

Abstract *The batch evaluation of information retrieval systems typically makes use of a testbed consisting of a collection of documents, a set of queries, and for each query, a set of judgements indicating which documents are relevant. This paper presents a probabilistic model for predicting IR system rankings in a batch experiment when using document relevance assessments from different judges, using the precision-at-n family of metrics. In particular, if a new judge agrees with the original judge with an agreement rate of α, then a probability distribution of the difference between the P@n scores of the two systems is derived in terms of α.*

We then examine how the model could be used to predict system performance based on user evaluation of two IR systems, given a previous batch assessment of the two systems together with a measure of the agreement between the users and the judges used to generate the original batch relevance judgements. From the analysis of data collected in previous user experiments, it can be seen that simple agreement (α) between users varies widely between search tasks and information needs. A practical choice of parameters for the model from the available data is therefore difficult. We conclude that gathering agreement rates from users of a live search system requires careful consideration of topic and task effects.

Keywords Information retrieval; Evaluation; User studies

1 Introduction

To test whether one information retrieval system is better than another, researchers either adopt the Cranfield methodology of *batch assessment*, or test their systems with humans in a *user experiment*. The batch assessment methodology requires a collection of documents, a set of queries, and, for each query, a judgment on some or all of the documents indicating whether they are relevant to that query or not. Assessing systems, therefore, is a matter of running each query to get a ranked list of documents, noting which is relevant or not according to the relevance judgements, and summarising the ranked list of relevance values into an overall performance score. The alternative approach requires a group of human users, the designing of a suitable experiment that controls for any biases you may wish to exclude (for example, education or computer literacy), defining an outcome metric (for example, time taken to find a useful answer document), and then measuring how users perform with different retrieval systems.

The batch method is by far the cheapest, easiest, and more repeatable of the two methodologies, and as such has dominated IR research for the last three decades. Recently, however, a series of papers has shown that the two methodologies do not necessarily reach the same conclusions regarding relative system performance. That is, if batch experiments show system A to be better than system B, user experiments may show there is no difference between the systems [1, 2, 5, 6, 7, 11, 13], or that system B is superior [12].

Our recent work has focussed on trying to quantify and rectify this seeming mismatch between the two experimental approaches [9, 10]. A key potential source of mismatch is the different relevance criteria of the judges used to construct the "ground truth" batch judgements, and the users in the user based experiment. Determining the relevance of a document to a query is a complex, multi-faceted task [4]. It often depends on the reason that the relevance judgement is being made, a *task effect*; the query itself, a *topic effect*; and of course the person making the judgement, a *judge effect*. There are many other factors that influence human judgement in general, including motivational biases, preconceptions, salience and availability, and perseverance [8]; these may all have additional effects on the criteria that judges use to decide if a document is relevant or not.

In this paper we develop a probabilistic model of agreement between relevance judges, and derive how this is expected to affect the results of a batch-based evaluation of IR system performance. We then investigate how agreement values could be obtained from a user study, so that the model might be used to transfer the outcomes from a batch experiment to a new user

Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia, 4 December 2009.

population. Analysis of the data from the user experiments shows that agreement between users is subject to substantial variation from both task and topic effects.

2 Preliminaries

Let a batch evaluation testbed consist of: a set of documents, $D = \{d_1, \ldots, d_{|D|}\}$; N queries, $Q = \{q_1, \ldots, q_N\}$; and for each query-document pair a relevance judgement

$$R(d_i, q) = \begin{cases} 1, & \text{document } d_i \text{ is relevant to query } q, \\ 0, & \text{otherwise.} \end{cases}$$

A *system*, returns a ranked list of m documents $[d_{i_1}, \ldots, d_{i_m}]$ for query q, which are mapped to a vector of relevance judgements in retrieved order $J = [R(d_{i_1}, q), \ldots, R(d_{i_m}, q)]$.

Judgement vectors can be reduced to a single score in various ways, and many different performance metrics have been proposed for representing the overall performance of IR systems. In this paper we use the precision-at-n documents metric, usually written P@n, which is the proportion of the top n documents of the list that are relevant. Formally,

$$\text{P@}n = \frac{1}{n} \sum_{i=1}^{n} J[i].$$

The *score* for a system is the mean P@n over all queries in the corpus. A system with a statistically significantly higher score than another is defined to be superior in the batch mode of system comparison, and is assumed to be superior in the user mode of comparison. This has been shown to be the case for P@1 in user experiments when the measured outcome is "time to save first relevant document" [9], and "satisfaction" [7], and for the tasks and users employed in those studies.

For example, if System B has a score of P@3=0.33, then on average only 1 of the top 3 documents is relevant, while if System A has a score of P@3=1.0, then the top 3 documents are always relevant for all test queries. It is implicitly assumed in IR experimentation that System A is superior to System B.

3 Modelling changes in judges

Using a testbed such as those from the Text REtrieval Conference [16] will yield system rankings that should be comparable with other experiments based on different queries with the same collection [15] or – with suitable standardisation – across different queries and collections [17]. That is, if System A is found to be statistically significantly better than System B when running a batch experiment, then this relationship should in general continue to hold for different queries, and collections. If, however, you kept the same set of documents and queries, but used an alternate relevance judge, so that $R(d, q)$ became $R'(d, q)$, then system rankings may alter. In particular, the P@n score for System B might increase, and the P@n score for System A might decrease, so that B becomes the superior system.

i	$J_A[i]$	$J_B[i]$	δ_i	$J'_A[i]$	$J'_B[i]$	δ'_i
1	1	0	1	0	0	0
2	1	0	1	1	1	0
3	0	0	0	0	1	-1
4	1	1	0	1	1	0
5	1	0	1	0	0	0
		$\Delta(5) = 3/5$			$\Delta'(5) = -1/5$	

Table 1: An example calculation of $\Delta(5)$ and $\Delta'(5)$ when System A has a P@5 value of 0.8 and 0.4 with different judgements, and System B has P@5 of 0.2 and 0.6 respectively.

If we assume that the new judge has some probability of agreeing with the judge used to build the original corpus (independently for any query-document pair), then we could derive a probability distribution of the new scores for System A and B. In turn, this can be used to derive a probability distribution on the difference between the two systems, and we can hypothesise about how transferable system rankings are between judges.

Definition 1 *Let J_A be the relevance vector given by System A for query q using corpus judgements $R(d, q)$, and J_B the relevance vector given by System B. For the same document lists, let J'_A and J'_B be the relevance vector given using judgements $R'(d, q)$ for System A and B respectively.*

We can now define the difference in P@n scores between the systems for query q using either set of relevance judgements, and then derive a probability distribution for that difference based on agreement probabilities between judges.

Definition 2 *For some ranked position $1 \leq i \leq n$, let $\delta_i = J_A[i] - J_B[i]$, and $\delta'_i = J'_A[i] - J'_B[i]$. Then the difference in P@n scores for the systems using either set of relevance judgements is given by $\Delta(n) = \sum_{i=1}^{n} \delta_i / n$ and $\Delta'(n) = \sum_{i=1}^{n} \delta'_i / n$ respectively.*

Table 1 shows an example of how $\Delta(5)$ and $\Delta'(5)$ is calculated. In this instance, using the second set of judgements has decreased System A's superiority to $\Delta'(5) = -0.2$, that is, System B is now apparently better than System A.

Without loss of generality, we will from now assume that System A has a higher P@n score than System B using the corpus judgements J_A and J_B. Thus we are interested in deriving a probability distribution for $\Delta'(n)$, and in particular the probability that $\Delta'(n) \geq 0$; that is, System A remains superior with a new set of judgements.

Definition 3 *Let α_0 be the probability that the new judge agrees with a $R(d, q) = 0$ judgement in the*

$J_A[i]$	$J_B[i]$	$J'_A[i]$	$J'_B[i]$	δ'_i	Probability	Probability$\times\delta'_i$
0	0	0	0	0	$\alpha_0\alpha_0$	0
0	0	0	1	-1	α_0(1-α_0)	-α_0(1-α_0)
0	0	1	0	1	(1-α_0)α_0	α_0(1-α_0)
0	0	1	1	0	(1-α_0)(1-α_0)	0
						$E_{00} = 0$
0	1	0	0	0	α_0(1-α_1)	0
0	1	0	1	-1	$\alpha_0\alpha_1$	-$\alpha_0\alpha_1$
0	1	1	0	1	(1-α_0)(1-α_1)	(1-α_0)(1-α_1)
0	1	1	1	0	(1-α_0)α_1	0
						$E_{01} = 1 - \alpha_0 - \alpha_1$
1	0	0	0	0	(1-α_1)α_0	0
1	0	0	1	-1	(1-α_1)(1-α_0)	-(1-α_0)(1-α_1)
1	0	1	0	1	$\alpha_1\alpha_0$	$\alpha_0\alpha_1$
1	0	1	1	0	α_1(1-α_0)	0
						$E_{10} = \alpha_0 + \alpha_1 - 1$
1	1	0	0	0	(1-α_1)(1-α_1)	0
1	1	0	1	-1	(1-α_1)α_1	-α_1(1-α_1)
1	1	1	0	1	α_1(1-α_1)	α_1(1-α_1)
1	1	1	1	0	$\alpha_1\alpha_1$	0
						$E_{11} = 0$

Table 2: All possible cases for judgement of a document in a ranked list at position i by the corpus and new judges, with their corresponding probabilities. For each possible pair of J_A and J_B values, the expected value of δ'_i, labelled E_x for each x, is computed as the sum of the four entries above it.

corpus, thus $R'(d,q) = 0$, and α_1 be the probability that the new judge agrees with a $R(d,q) = 1$ judgement in the corpus, hence $R'(d,q) = 1$.

For any rank i in the top n documents for a single query, the entries in the relevance vectors for System A and System B for that position is either: $J_A[i] = 0$ and $J_B[i] = 0$, both systems returned an irrelevant document in that position; $J_A[i] = 1$ and $J_B[i] = 1$, both system returned a relevant document in that position; and the two discriminating cases $J_A[i] = 1$ and $J_B[i] = 0$, or $J_A[i] = 0$ and $J_B[i] = 1$. Table 2 shows, for each of these four possible cases, the four possible relevance vector entries at a particular rank i that might result using different judgements ($J'_A[i]$ and $J'_B[i]$). In addition to the δ'_i values for each case, the probability of realising each combination is given in the second last column, which is the product of the appropriate agreement probabilities. For example, in the first row the probability of $J_A[i] = 0$ and $J'_A[i] = 0$ is α_0, and $J_B[i] = 0$ and $J'_B[i] = 0$ is also α_0, so total probability of that event is $\alpha_0\alpha_0$. In the second row, $J_A[i] = J'_A[i] = 0$, but $J_B[i] = 0$ is judged as $J'_B[i] = 1$ with probability $(1 - \alpha_0)$, so the total probability is $\alpha_0(1 - \alpha_0)$. The final column is summed for each of the four possible cases of $J_A[i]$ and $J_B[i]$ to give the expected value of δ'_i for that case, labelled E_{00}, E_{01}, E_{10}, and E_{11} respectively.

Definition 4 *For a given query q and Systems A and B, let c_{00} be the number of rank positions in the top n for query q where $J_A[i] = 0$ and $J_B[i] = 0$, and likewise for c_{10}, c_{01} and c_{11}. That is, $c_{xy} = |\{J_A[i] = x \text{ and } J_B[i] = y, 1 \le i \le n\}|$. Note, $\Delta(n) = (c_{10} - c_{01})/n$.*

For each position in a ranked list, $E_{J_A[i]J_B[i]}$ gives the expected value of δ'_i, and so the expectation of $\Delta'(n)$ can be calculated as:

$$\begin{aligned} E[\Delta'(n)] &= E\left[\sum_{i=1}^{n} \delta'_i/n\right] \\ &= (c_{00}E_{00} + c_{01}E_{01} + c_{10}E_{10} \\ &\quad + c_{11}E_{11})/n \\ &= (1 - \alpha_0 - \alpha_1)(c_{01} - c_{10})/n \\ &= (\alpha_0 + \alpha_1 - 1)\Delta(n) \end{aligned} \quad (1)$$

Intuitively this makes sense. If new judges agree perfectly with the corpus judges, then $\alpha_0 = \alpha_1 = 1$, then $E[\Delta'(n)] = \Delta(n)$: there is no expected difference in the system's scores with either judgement set. If new judges disagree completely with the corpus judges, then $\alpha_0 = \alpha_1 = 0$, then $E[\Delta'(n)] = -\Delta(n)$: that is, the expected system scores are the reverse of the original.

We can also compute the variance of $\Delta'(n)$. Recall that $\text{Var}(X) = E(X^2) - E(X)^2$ by definition, so:

$$
\begin{aligned}
& \text{Var}(\Delta'(n)) \\
= \; & \text{Var}\left(\sum_{i=1}^{n} \delta_i'/n\right) \\
= \; & \sum_{i=1}^{n} \text{Var}(\delta_i')/n^2 \\
= \; & \frac{1}{n^2}\sum_{i=1}^{n}\left(E[(\delta_i')^2] - E[\delta_i']^2\right) \\
= \; & (c_{00}(2\alpha_0(1-\alpha_0)) \\
& +c_{11}(2\alpha_1(1-\alpha_1)) \\
& +(c_{01}+c_{10})(1-\alpha_0-\alpha_1+2\alpha_0\alpha_1) \\
& -(1-\alpha_0-\alpha_1)^2(c_{01}-c_{10})^2)/n^2 \qquad (2)
\end{aligned}
$$

Equations 1 and 2 are for a single query, q, but are easily extended to a score computed over a set of N queries because the P@n metric assigns equal weight to all ranked positions. That is, computing the mean P@n value over the top n documents retrieved for N queries is the same as computing P@Nn for a concatenation of the N $J[1..n]$ relevance vectors for each query. If we use the notation J_i to represent the relevance vector J for query i, and $J_S = J_1[1..n]J_2[1..n]..J_N[1..n]$ to represent the concatenation of the first n elements of all J_i's, then:

$$
\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}(\text{P@}n \text{ of } J_i) &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n}\sum_{j=1}^{n} J_i[j] \\
&= \frac{1}{Nn}\sum_{k=1}^{nN} J_S[k] \\
&= \text{P@}Nn \text{ of } J_S.
\end{aligned}
$$

Henceforth we will limit our discussions to the single query case for notational convenience.

Equation 2 contains c_{xy} terms, which will alter depending on system, query and judgements. However, if we fix n, or assume the maximum possible separation between systems on the corpus, the equations can be simplified to something immediately useful.

3.1 The P@1 case

When considering P@1, the expression for $\text{Var}(\Delta'(n))$ simplifies to something manageable. As we are interested in the case where System A is better than System B on query q using the corpus judgements, then P@1=1 for System A and for System B, P@1=0. Hence $c_{00} = c_{11} = c_{01} = 0$, $c_{10} = 1$, $n = \Delta(n) = 1$, and

$$
\begin{aligned}
E[\Delta'(n)] &= \alpha_0 + \alpha_1 - 1 \\
\text{Var}(\Delta'(n)) &= \alpha_0 + \alpha_1 - \alpha_0^2 - \alpha_1^2.
\end{aligned}
$$

Assuming $\Delta'(n)$ is normally distributed with mean $E[\Delta'(n)]$ and a standard deviation of $\sqrt{\text{Var}(\Delta'(n))}$, then we can compute $Pr[\Delta'(n) \geq 0]$ which is shown in Figure 1. To be more than 50% confident that a new set of judgements on the corpus will keep System A as superior with the P@1 metric, the sum of α_0 and α_1 must be larger than 1 (approximately). To be 95% confident that System A will remain superior, both agreement probabilities must be over 80%.

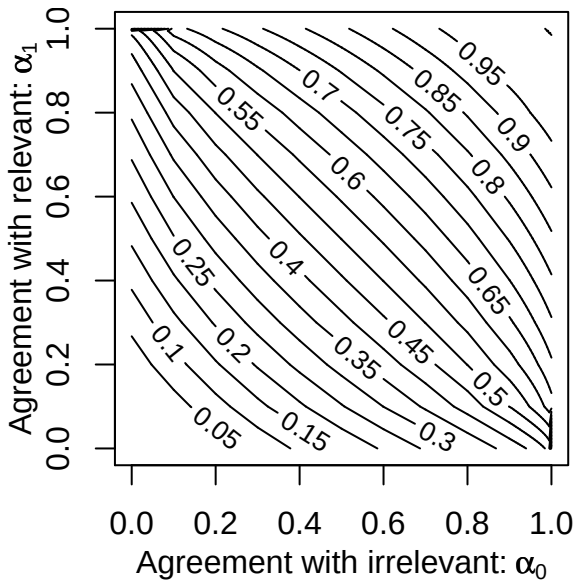

Figure 1: Contour plot of the probability of $\Delta'(1)$ exceeding zero (hence System A remaining superior) with the P@1 score, when the corpus is re-judged by a judge that agrees α_0 and α_1 proportion of the time with the original judge's 0 and 1 judgements, respectively.

3.2 The extreme case

Just as for the P@1 case, assuming P@n=1 for System A and P@n=0 for System B allows simplification of Equations 1 and 2 as all of c_{00}, c_{11} and c_{01} are 0, and $c_{10} = n$.

Thus

$$
\begin{aligned}
E[\Delta'(n)] &= \alpha_0 + \alpha_1 - 1 \\
\text{Var}(\Delta'(n)) &= ((1-\alpha_0-\alpha_1+2\alpha_0\alpha_1) \\
&\quad -(\alpha_0+\alpha_1-1)^2)/n
\end{aligned}
$$

If we assume that $\alpha_0 = \alpha_1 = \alpha$, then we can plot $E[\Delta(n)]$ and a 95% confidence interval as $\pm 1.96\sqrt{\text{Var}(\Delta'(n))}$ for different n values. This is shown in Figure 2.

To be 95% sure that System A remains superior with new judgements, agreement must be at least 90% for P@1 (intersection of dark grey ellipse and the 0 line), 75% for P@5 (intersection of medium grey ellipse and the 0 line), and 70% for P@10 (intersection of light grey ellipse and the 0 line).

3.3 Other cases

It is possible to simplify Equations 1 and 2 for other values of n where System A and System B are not separated extremely, that is, when the gap between System A and System B is less than one: $\Delta(n) < 1$. The technique involves labelling each possible combination of $J_A[i]$ and $J_B[i]$ for all i, but is omitted from this

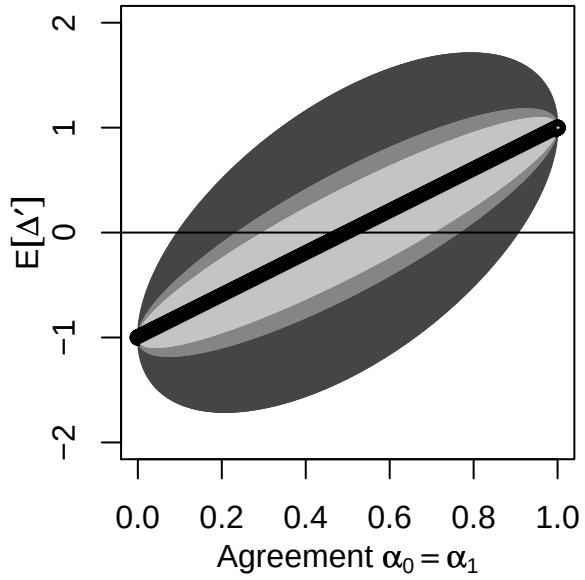

Figure 2: Expected $\Delta'(n)$ values (black) and 95% confidence limits for $n = 1$ (dark grey), $n = 5$ (medium grey) and $n = 10$ (light grey) assuming P@1=1 for System A and P@1=0 for System B.

paper as we concentrate on the P@1 metric in our user studies.

4 Practical considerations

In this section of the paper we turn our attention to an investigation of the likely values α_0 and α_1 when users conduct a web-based search task. In particular, we examine data from one of our previous user studies that involved both document judgements and search-and-click judgements, and see if α values are stable across different topics and tasks for a given pair of users.

4.1 User experiment

Participants for our user study were recruited form RMIT University. All were postgraduate or undergraduate students studying for degrees in computer science and information technology. As a result, most were very familiar with searching for information on the web; in a pre-experiment questionnaire the average user indicated that they search "once or more a day". Experiments were carried out in compliance with the RMIT University Human Research Ethics Committee. 40 users participated in the study; however, three were unable to complete the full experiments, and are therefore excluded from the analysis.

Participants were asked to carry out two tasks: a judging task, and a search task. For both, documents and topics were sourced from the TREC GOV2 collection, a crawl of 426 Gb of data from the `.gov` domain from 2004 [3].

Judging task: For the first task, participants were asked to imagine that they are writing a report, based on a provided information need, and to mark documents that were presented as relevant or not relevant for inclusion in the report. Participants were asked to carry out this task for three TREC topics (numbers 707, 770 and 771); the description and narrative fields of the topics were displayed to users as information needs. Participants were therefore making binary decisions about relevance, when presented with documents that had previously been judged by TREC assessors on a three-point scale (not relevant; relevant; and highly relevant). There was no time constraint for making decisions for the judging task. However, it became clear that carrying out the task for all three topics resulted in severe fatigue effects. The third topic completed by each user is therefore removed from the analysis.

Searching task: Participants also carried out a searching task. Here, when presented with an information need, users were asked to search for and identify a relevant answer document as quickly as possible. Users could enter a single query to a search system, designed to be similar in appearance to popular commercial search engines such as Google, Yahoo! or Bing. Unknown to the user, for each topic they were assigned to a system of a particular quality; that is, the system would return a ranked answer list with a pre-determined P@1 level. For this task, 24 informational topics were chosen from TREC topics 700–850 (topics developed for use with the GOV2 collection). To construct the P@1 controlled lists, judged documents were selected from the two highest-performing runs submitted to the TREC terabyte track in 2004, 2005 and 2006. That is, all documents used in the lists could plausibly be returned in response to the topics by a modern information retrieval system.

After being presented with a search results list, a user could select a document for viewing. They could then take one of two actions: save the document as a relevant answer; or close the document, and return to the results list. In the analysis below, these actions are taken as judgements of the relevance or non-relevance of the document, respectively.

Note that the user studies were not explicitly designed to answer the questions raised in this paper; rather we are retrospectively analysing the data to get insights into likely values of α_0 and α_1. Full details of the user studies are available in previous papers [9, 10].

4.2 Agreement on the judging task

Figure 3 shows the distribution of agreement values between all pairs of users for the judging task. As agreement is not symmetrical [14], each user pair is counted twice, usually with different values. As can be seen, agreement varies anywhere from 100% down to 7.7% for users 14 and 11 on α_0.

Perhaps of more interest is the difference in agreement for any user pair that judged the same two topics. Figure 4 shows that on any two topics, both α_0 and

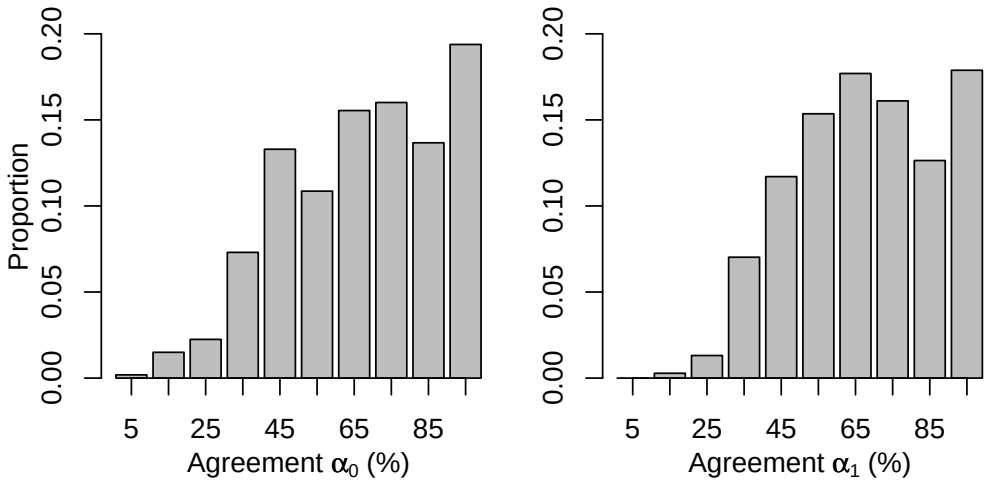

Figure 3: Distribution of agreement amongst all pairs of users on the judging task.

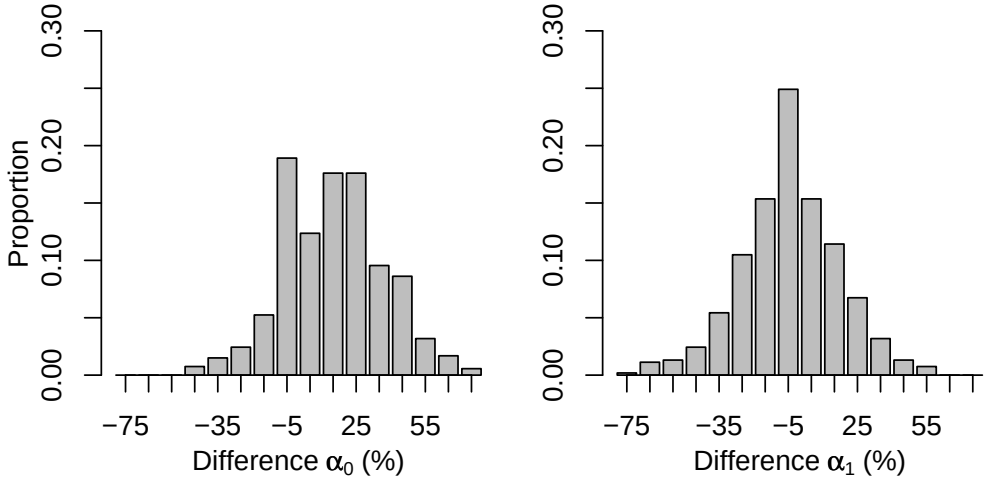

Figure 4: Distribution of the difference in agreement amongst pairs of users on the judging task.

α_1 can vary widely in the judging task. This makes it difficult to choose a representative agreement value for any pair of judges. Note that as we had to remove the third topic judged for each user from the data set, not all pairs of users completed the same two topics. In total 534 of the 1369 pairs are included.

4.3 Agreement on the search task

Figure 5 shows the distribution of agreement values between all pairs of users for the searching task. Here we have taken the event where a user selected a document from the ranked list but did not save it as an "irrelevant" judgement, while the selection and explicit saving of an item is taken as a "relevant" judgement. For any pair of users, we computed α_0 and α_1 over all topic-document pairs that both users selected from the ranked lists for viewing. We only included pairs where at least 6 topic-document pairs were judged as relevant and irrelevant by the first user in the pair, giving 758 user pairs. Again, agreement is not symmetric, and so each pair of users is counted twice, typically with different values. As can be seen, the distribution of agreement values is similar to those for the judging task.

4.4 Agreement across tasks

Figure 6 shows the distribution of the difference in α_0 and α_1 for pairs of users between the searching and judging tasks. Again, the difference across tasks can be large, making it difficult to choose a representative agreement value for any pair of judges/users.

Figure 7 plots each user pair that has an agreement value for both tasks. As is apparent, there is no guarantee that if a pair of users did not agree in the judging task, they will not agree in the search task, and vice versa.

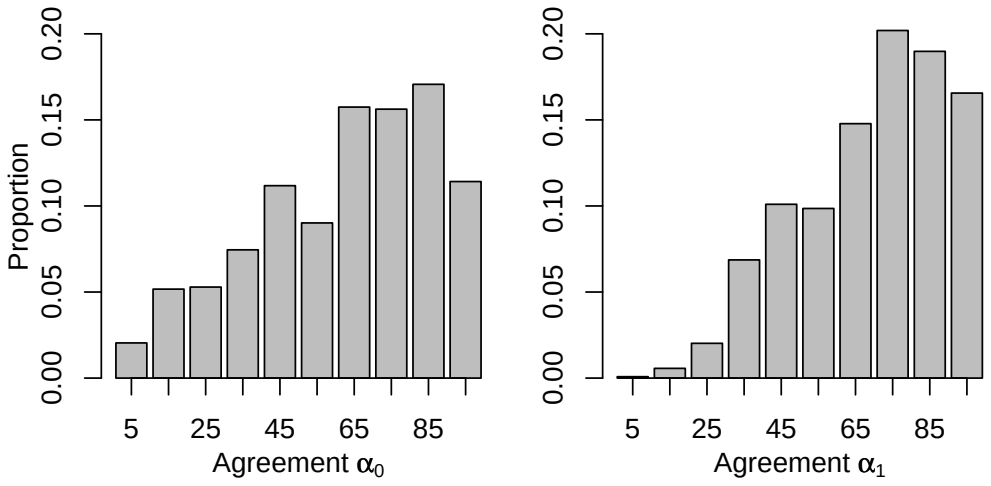

Figure 5: Distribution of agreement amongst all pairs of users on the searching task.

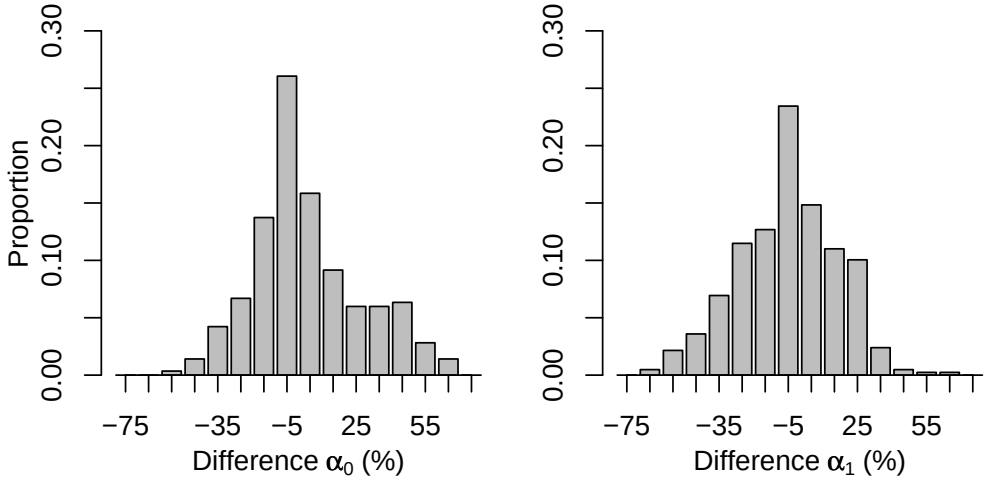

Figure 6: Distribution of the difference in agreement amongst pairs of users on the searching task and judging task.

5 Conclusions

We have presented a simple probabilistic model based on agreement between judges that can predict the effect that altering judges will have on system performance as measured through a batch evaluation experiment. When evaluating performance with P@1, for example, to be 95% confident that one system will remain superior to a second after judges are changed, the agreement between relevance assessments of the judges must be at least 80%.

The model can also be used to assist in selecting metrics. For example, for the P@n family of metrics, it can be seen that the larger the value of n (that is, the more information from the result list that is considered), the lower the required level of agreement between judges to remain confident that the relative system performance will not change. In this paper we have concentrated on the P@n metrics; in future work we plan to extend the approach to other metrics.

Examining the agreement values in one of our user studies has revealed large topic and task effects. That is, for any pair of users, their agreement may alter on different topics or tasks by over 50%. Thus, applying the model presented in Section 3 to predict the effect of changing judges on a corpus requires more sophisticated measuring of α_0 and α_1 than was possible with our available user data. In future work, we plan to investigate controlled experiments for gathering representative agreement values between different users of retrieval systems.

References

[1] Azzah Al-Maskari, Mark Sanderson and Paul Clough. The relationship between IR effectiveness measures and user satisfaction. In *Proceedings of the ACM SIGIR*

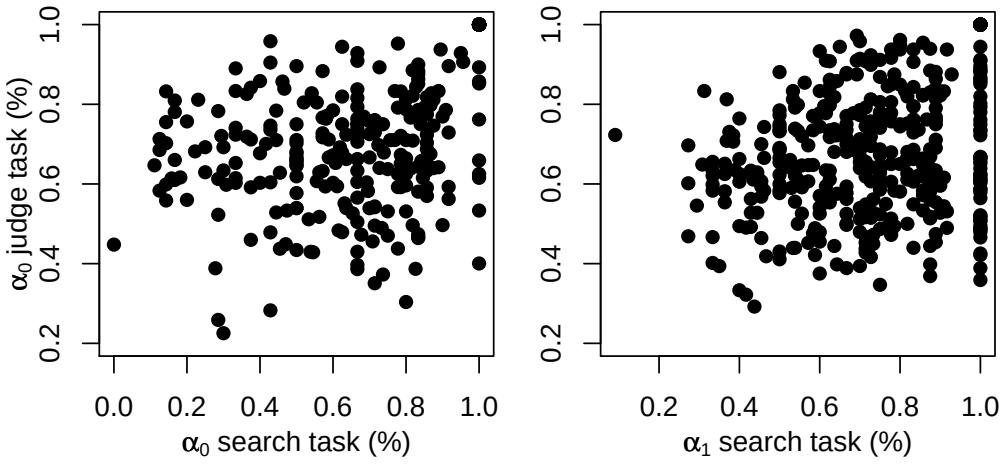

Figure 7: The agreement on each task for each pair of users.

International Conference on Research and Development in Information Retrieval, pages 773–774, Amsterdam, Netherlands, 2007.

[2] James Allan, Ben Carterette and Joshua Lewis. When will information retrieval be "good enough"? In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 433–440, Salvador, Brazil, 2005.

[3] Stefan Büttcher, Charles Clarke and Ian Soboroff. The TREC 2006 terabyte track. In *The Fifteenth Text REtrieval Conference (TREC 2006)*, Gaithersburg, MD, 2007. National Institute of Standards and Technology.

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

[5] William Hersh, Andrew Turpin, Susan Price, Benjamin Chan, Dale Kraemer, Lynetta Sacherek and Daniel Olson. Do batch and user evaluations give the same results? In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 17–24, Athens, Greece, 2000.

[6] Scott B. Huffman and Michael Hochster. How well does result relevance predict session satisfaction? In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 567–574, Amsterdam, Netherlands, 2007.

[7] Diane Kelly, Xin Fu and Chirag 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.

[8] Arie Kruglanski and Icek Ajzen. Bias and error in human judgement. *European Journal of Social Psychology*, Volume 13, pages 1–44, 1983.

[9] Falk Scholer and Andrew Turpin. Metric and relevance mismatch in retrieval evaluation. In *The Fifth Asia Information Retrieval Symposium (AIRS 2009)*, Sapporo, Japan, 2009. To appear.

[10] Falk Scholer, Andrew Turpin and Mingfang Wu. Measuring user relevance criteria. In *The Second International Workshop on Evaluating Information Access (EVIA 2008)*, pages 47–56, Tokyo, Japan, 2008.

[11] Catherine Smith and Paul Kantor. User adaptation: good results from poor systems. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 147–154, Singapore, Singapore, 2008.

[12] Andrew Turpin and William Hersh. Why batch and user evaluations do not give the same results. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 225–231, New Orleans, LA, 2001.

[13] Andrew Turpin and Falk Scholer. User performance versus precision measures for simple web search tasks. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 11–18, Seattle, WA, 2006.

[14] Alexander von Eye and Eun Young Mun. *Analyzing Rater Agreement: Manifest Variable Methods*. Lawrence Erlbaum Associates, 2004.

[15] Ellen M. Voorhees and Chris Buckley. The effect of topic set size on retrieval experiment error. In Micheline Beaulieu, Ricardo Baeza-Yates, Sung Hyon Myaeng and Karlervo Järvelin (editors), *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 316–323, Tampere, Finland, 2002.

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

[17] William Webber, Alistair Moffat and Justin Zobel. Score standardization for inter-collection comparison of retrieval systems. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 51–58, Singapore, Singapore, 2008.