Metric and Relevance Mismatch in Retrieval Evaluation

Falk Scholer and Andrew Turpin

School of Computer Science and IT, RMIT University
GPO Box 2476v, Melbourne, Australia
{falk.scholer,andrew.turpin}@rmit.edu.au

Abstract. Recent investigations of search performance have shown that, even when presented with two systems that are superior and inferior based on a Cranfield-style batch experiment, real users may perform equally well with either system. In this paper, we explore how these evaluation paradigms may be reconciled. First, we investigate the DCG@1 and P@1 metrics, and their relationship with user performance on a common web search task. Our results show that batch experiment predictions based on P@1 or DCG@1 translate directly to user search effectiveness. However, marginally relevant documents are not strongly differentiable from non-relevant documents. Therefore, when folding multiple relevance levels into a binary scale, marginally relevant documents should be grouped with non-relevant documents, rather than with highly relevant documents, as is currently done in standard IR evaluations.

We then investigate relevance mismatch, classifying users based on relevance profiles, the likelihood with which they will judge documents of different relevance levels to be useful. When relevance profiles can be estimated well, this classification scheme can offer further insight into the transferability of batch results to real user search tasks.

1 Introduction

Information retrieval (IR) experiments based on the Cranfield methodology measure system performance using a set of queries and a test collection. The queries are run over the collection using a search system, and for each document that is returned, a human judge decides whether the document is relevant to the query, or not. The overall utility of the search system is then computed using a metric that aggregates the relevance judgements for documents in ranked lists returned by the system. In this batch evaluation approach, different search systems are compared based on how well they score on such metrics. For example, many papers report IR system comparisons using the TREC document collections, topics and judgements, using Mean Average Precision (MAP) or Precision at 10 documents retrieved (P@10) as the metric [23].

An alternate way to evaluate systems is to take a group of human users and ask them to perform search tasks with different systems, comparing outcome measures such as time to complete a task, success or failure on a task, or subjective measures like user satisfaction. Previous studies [1,2,8,9,12,18,19] have

G.G. Lee et al. (Eds.): AIRS 2009, LNCS 5839, pp. 50–62, 2009.

shown that attempting to transfer results from batch experiments to real users is difficult. That is, the systems rated as superior in the batch experiments may in fact not assist users in performing their tasks more quickly or more accurately than the systems that are rated more poorly in the batch experiments.

In this paper, we explore ways in which these two experimental paradigms may be reconciled. There are many possible causes for this seeming mismatch between batch and user-based experimental outcomes. We investigate two reasons using controlled batch and user experiments.

Mismatching metrics. It is possible that the metric used in a batch experiment to show that System A is superior to System B does not reflect the user task for which these systems will be employed. For example, if a batch experiment uses the MAP metric, which contains a recall component, but the user task is solely precision based, such as finding a single answer to a simple question, then differences between systems in the batch experiment may be meaningless in the user domain. On the other hand, if the batch experiment used a metric such as Precision at one document returned (P@1) or at three documents returned (P@3), then it is perhaps more likely that the batch results would carry over into the user domain. For example, Turpin and Scholer [19] used the MAP metric to choose superior systems, but then employed those systems on a precision-based user task and found that they did not outperform the inferior systems. When they re-analysed their data to choose systems based on the P@1 metric, it suggested that users performed better with the superior system. However, the analysis of P@1 was inconclusive because of the small number of systems for the non-relevant category of this metric. Motivated by this finding, we explicitly examine possible metric mismatch by using P@1 in our batch experiments, and a precision-based user outcome measure. We also extend this analysis to incorporate multiple levels of relevance, factoring in differences between non-relevant, relevant, and highly relevant documents.

Mismatching relevance profiles. Batch system results are based on relevance judgements assigned to documents by human assessors. However, it is possible that relevance judgements used in the batch experiments are made using different criteria, or on a different scale (whether perceptual or actual), than judgements that are made in a user study. For example, in this study we use TREC documents that are judged on a three-point scale: non-relevant (0), relevant (1), and highly relevant (2). The TREC judging criteria define level zero as being applicable where no “part of the document contains information which the assessor would include in a report on the topic”; while the distinction between level one and two was “left to the individual assessors to determine” [7]. If subjects in a user study receive identical instructions to the judges in the batch experiment, and carry out their evaluation in as similar an environment as possible, there is still scope for individuals to decide their own threshold on what information they would “include in a report”, and to distinguish between the two categories of relevance. Even in the highly controlled TREC judging environment, the overlap between the relevance judgements of assessors is on average only about 45% [22], indicating that thresholds between relevance categories can differ even within relatively homogeneous populations where identical judging instructions are given. Therefore relevance mismatch, where users and batch judges have different expectations and preferences for documents of different relevance levels, may lead to conflicting results between batch and user results. We investigate the impact of relevance mismatch based on the *split agreement* approach [14], where users are classified into groups based on their responses to documents of different relevance levels.

These two possible explanations for differences between batch and user experiments are investigated through a user study. In Section 2 we survey related background work on experimental evaluation in IR. Our experimental methodology, including details of the user-based searching task, is explained in Section 3. Results are presented and discussed in Section 4, with conclusions and further work being considered in Section 5.

2 Background

The Cranfield paradigm of information retrieval evaluation involves using a search system to run a set of queries on a fixed collection of documents. For each potential answer that the search system returns, a human is required to judge the relevance of the particular document for the current query. This is the dominant framework for experimental IR, and is used, for example, in the ongoing series of Text REtrieval Conferences (TREC). TREC provides standard collections, queries, and relevance judgements so that the performance of different IR systems can be compared using common testbeds [23]. In TREC, queries are derived from topics that represent user information needs: topics consists of a *title* field (a small number of keywords, representative of what a user might type into a web search engine), a *description* (a longer statement of the topic, usually a single sentence), and a *narrative* (a short paragraph specifying further requirements) [6].

Based on the system search result lists and relevance judgements, different system performance metrics can be calculated. Many metrics that have been proposed in the literature focus on precision, which is the number of relevant documents that the search system has found as a proportion of the total number of documents that the system has returned. *Average precision* (AP) is calculated as the mean of the precision at each relevant item that occurs in a result list for a single query. Relevant documents that are not returned by the system contribute a precision of zero; this metric thus has a recall component, since the system is penalised for missing answers. Across a set of queries, the mean average precision (MAP) provides a single number that summarizes search performance, reflecting both the precision and the recall of the system [5].

Another widely-used class of performance metrics is the precision of a system at a particular cutoff point N in the search results list. For example, P@1 evaluates a system based on the relevance of the first item in the result list, while P@10 calculates the precision over the first 10 results. These metrics are popular for evaluating web search tasks, since users typically focus on results that occur early in the ranked list [17]. Analysis by Buckley and Voorhees has indicated that these P@N metrics require a relatively larger number of test queries, compared to other metrics such as MAP, in order to give stable results for the evaluation of batch experiments [4].

The most commonly used IR system performance metrics, such as those presented previously, treat relevance as a binary criterion: a document is either relevant, or it is not. Even where documents may have been judged on a multiple-level relevance scale, these levels are typically folded together into a binary classification before the metrics are calculated. However, studies of multiple levels of relevance have indicated that the traditional binary relevance assumption may not be appropriate where actual users of search systems are concerned [16,21]. In the TREC evaluation framework, the criterion for relevance states that if the document includes any reference to the topic, it should be counted as being relevant. This includes documents that are only *marginally* relevant, where the document does not contain information other than that contained in the topic description; in other words, these documents are largely useless from a user's perspective. Investigating the ability of users to judge documents of different relevance levels, Vakkari and Sormunen concluded that the likelihood of identifying highly relevant documents is much higher than for marginally relevant ones [21]. Further, analysis of 38 topics from TREC-7 and 8 by Sormunen showed that around 50% of documents that were judged as relevant under the TREC binary criterion were of this marginal category [16]. We investigate the effect of accounting for different levels of relevance has on the results of user-based and batch retrieval experiments.

The *cumulative gain* (CG) family of retrieval metrics are based on the idea that the relevance of documents is not equal: the usefulness to a user will depend on the level of relevance of an item [11]. This allows multiple levels of relevance to be incorporated in system evaluation, unlike the previously discussed metrics which assume a binary relevance scale. The CG values, where more highly relevant documents are rewarded by adding more to the overall performance score, can then be *discounted* (DCG) so that the further a document is from the top of a ranked list, the more heavily its relevance score is adjusted. In this paper, we investigate DCG@1 as a multiple-relevance level alternative to P@1. Since discounting is usually not applied at the first rank of the answer list, CG@1 and DCG@1 are equivalent.

The Cranfield paradigm of IR evaluation makes a number of simplifying assumptions about users: essentially, users and real search tasks are removed from the evaluation process, with both information needs and relevance being reduced to static components of the analysis. While this allows for repeatability of experiments, and the controlled evaluation of retrieval algorithms, it is widely acknowledged that these assumptions are significant simplifications of the actual retrieval process [10]. A number of studies have therefore investigated the relationship between system-centric retrieval performance metrics and the performance of users engaged in a range of different search tasks, which we briefly survey here. A relationship between the ability of users to find answer facets and high changes in the level of the bpref evaluation metric was found by Allan et al. [2]. Investigations by Hersh and Turpin found no relationship between MAP and user performance on an instance recall task [8], or a question answering task [18]. The relationship between simple web search tasks and MAP was investigated by Turpin and Scholer [19]; no relationship was found with a precision-oriented task, but a weak relationship was observed with a recall-oriented task.

Other recent studies have considered the relationship between result relevance and user satisfaction. Experiments by Huffman and Hochster showed that system performance measured by DCG@3 was related to user satisfaction for informational searches [9]; user satisfaction was measured by asking subjects to rate their overall search experience on a seven-point scale. Al-Maskari et al. [1] compared the precision and various cumulative-gain metrics of search results with user satisfaction. Here, users rated their satisfaction based on the accuracy, coverage and ranking of results. A high correlation was found between satisfaction and both the precision and CG metrics, while the correlation with nDCG was low. In a series of carefully controlled experiments, Kelly et al. [12] demonstrate a strong correlation between precision and user satisfaction; ranking also influenced user ratings, but to a lesser extent. In this paper, instead of using self-reported measures of satisfaction, we investigate user performance based on success in completing a simple search task, measuring the time taken to find a relevant document.

To construct user relevance profiles, Scholer, Turpin and Wu proposed the *split agreement* approach [14]. Here, users are analysed based on their rate of agreement when presented with documents at different TREC relevance levels. Users can deviate from TREC-like relevance behaviour in two ways: *generous* users have lower criteria for relevance than TREC judges, and are often satisfied even with non-relevant (level 0) documents. Conversely, *parsimonious* users have stricter relevance criteria than TREC judges, and are usually satisfied only with a highly relevant (level 2) document. Users who are *TREC-like* follow the assumed batch relevance profile, generally discarding level 0 documents, but liking level 1 and 2 documents.

Relevance profiles are established through repeated presentation of documents with different TREC relevance levels (unknown to the user). For each presented document, the user is asked to indicate whether they find the document to be relevant for a specified information need, or not. Across many presentations of documents, a response proportion can thus be calculated for each TREC relevance level. For example, a particular user may judge level 0, 1 and 2 documents to be relevant 6%, 63% and 94% of the time, respectively.

User classes are based on these proportions. Specifically, a generous user is defined as someone who judges level 0 documents to be relevant *more* than 50% of the time. A parsimonious user, on the other hand, judges level 1 documents to be relevant *less* than 50% of the time [14]. To investigate relevance mismatch, we attempt to classify users based on their relevance preferences, aiming to establish for each user whether their relevance profile is similar to that of TREC judges.

3 Experimental Methodology

This study investigates the relationship between the P@1 and DCG@1 system performance metrics and user performance on a web search task, and how this is affected by user perceptions of relevance. We use TREC data for the basis of our batch experiments, and a user study to collect data for a searching task.

Users and document collection. 40 experimental subjects were recruited from RMIT University by advertising on newsgroups and notice-boards. All subjects were required to complete entry questionnaires. Participants were university students undertaking undergraduate or postgraduate studies in computer science and information technology, and most were very familiar with online searching (the median response for searching frequency was that searches are conducted "once or more a day"). Subjects were from a variety of cultural backgrounds, but all had a reasonable grasp of the English language (a requirement for studying at RMIT University). Experiments were carried out in accordance with the guidelines of RMIT University Human Research Ethics Committee. Three of the 40 user study participants were unable to carry out all required aspects of the experiments, and are excluded from the analysis below.

The documents used for the searching task are from the TREC GOV2 collection, a 426 Gb crawl of the US `.gov` domain carried out in 2004 [6]. This collection was used for the TREC Terabyte tracks in 2004–2006, and has 150 associated search topics and corresponding relevance judgements, made by NIST assessors. The relevance judgements are on a three-level ordinal scale: not relevant (0); relevant (1); and highly relevant (2). According to the standard TREC judging approach, if any part of the document contains information that the assessor would include in a report on the topic, it should be judged relevant [7]. That is, relevant documents will include those that are only of marginal value, containing little or no information beyond what is already included in the topic statement.

Search systems. To investigate the relationship between user search performance and system performance as measured by a batch metric, we mimic batch experimental results by constructing ranked lists using the known TREC relevance levels of documents to achieve a given level of the performance metric under investigation. A set of ranked lists at a given level can be thought of as being generated by a search *system* that is engineered to always produce ranked lists that achieve a particular level of the metric, for any topic.

Given that the TREC relevance judgements have three levels, there are thus three possible systems for the DCG@1 metric, namely lists starting with a document of relevance level 0, 1, or 2. For P@1, a binary metric, these relevance levels are folded together: either level 0 compared to combined levels 1 and 2;

or combined levels 0 and 1 compared to level 2. To reduce variation, all system lists had identical TREC relevance scores assigned after the first position. The document relevance level allocations for complete system lists were

$$X, 1, 1, 1, 0, 2, 0, 0, 1, 0$$

where $X \in \{0, 1, 2\}$. Lists were constructed to a depth of 10 documents.

For the search task, 24 topics were chosen from TREC topics 700-850; the constraint for topic selection was that each topic must have the required number of documents at each relevance level to allow the construction of the appropriate lists. Documents were assigned to lists by relevance level, with candidate documents being drawn from the top 50 documents from the two runs with highest MAP scores submitted to the Terabyte track for 2004, 2005 and 2006; that is, they are documents that would feasibly be returned in response to the topic by a modern search system. Only documents of type "text/html" were retained, with other content types being discarded. Similarly, documents smaller than 750 bytes or larger than 100,000 bytes were discarded.

Search task and user interface. Users were asked to carry out a *precision-based* search task: to quickly find useful information about a topic. This type of search is common on the web, and can be considered to be a simple instance of the informational search categories identified by Rose and Levinson [13]. As a performance outcome, we measure the amount of time that a user needs to complete the task.

Specifically, the search scenario is that of a user being asked to find useful information about a topic:

> "Imagine that your boss has come running into the room and urgently needs information. He gives you a very quick topic description, and you have only a few minutes to find a document that is useful (that is, contains some information about the requested topic)."

The information needs were framed in a task-based scenario so as to ground them in a practical context; Borlund has demonstrated that searcher behaviour that is elicited through simulated search tasks may be similar to behaviour that is exhibited when engaged with real information needs [3].

A search session proceeded as follows. First, a subject was presented with an information need, comprised of the *narrative* and *description* fields of a TREC topic, at the top of the screen. Under the information need, a search interface was available. This was closely modelled on the search screens of popular Web search engines, and consisted of a text-box for the entry of search terms, together with a "search" button. After a user entered a query, they were presented with a results list of the required system level (that is, corresponding to one of the precision variants, as described previously). Users were not able to reformulate their queries.

Entries in the search results lists consisted of the document title, together with a short query-biased summary. The document summaries were generated following the approach of Turpin et al. [20], using the *title* field of the TREC topic as the query words. The document title was a hyperlink which, when clicked, opened the underlying document in a new window.

From the document window, in addition to being able to read the document, subjects were presented with two option buttons: “save”, to mark the document as relevant to the information need; and “cancel”, to close the document window and return to the search results list. Choosing to save a document brought up a confirmation dialogue box, which asked the subject to enter a brief description of why the document was considered to be relevant. After saving one document, the user is deemed to have completed that particular search task.

All interactions between users and the search system were written to a system log, including timestamps of when actions took place. Timings for the precision-based search task were calculated from when the user clicked the search button, until they chose to save their document.

Users were asked to carry out searches on the 24 topics three times, so that each topic would be completed with every system level. The experimental design ensured that users were presented with topics and systems in different orders, to account for possible biases and learning effects. Due to fatigue that was apparent in the last half hour of the user study, we only analyse the first 48 (out of 72) total searches for each user below. However, due to rotation in the experimental design, the results are balanced so that, across all searches, topic and system combinations were used an equal number of times.

4 Results

Based on the user study, we investigate whether system differences as shown by batch metrics that focus on the relevance of the top-ranked position in a search result list transfer successfully to an actual search task. We then examine relevance profiles, and whether these can help to explain the relationship between the two evaluation paradigms.

4.1 Comparing Batch and User Performance

To investigate our first hypothesis, that the P@1 and DCG@1 system performance metrics are closely matched to a precision-based user search task, we analyse the relationship between search system level and time taken to find a useful document. The mean and median times that a user needed to find a useful document with different systems are shown in Table 1. On average, the task time falls as the level of the system performance metric rises. A multifactorial analysis of variance (ANOVA) indicates that the effect of the different system levels is statistically significant ($p < 0.0001$). However, the time data from the user search task is truncated at zero, and so violates the normality assumption. Although ANOVA is generally robust, we therefore also analyse system effects using the Kruskal-Wallis test, a non-parametric alternative to ANOVA [15]; this supports the previous results, also showing a statistically significant effect for

Table 1. Average time (in seconds) for a user to save a relevant document using different systems

System	Mean	Median
0	117.86	89.55
1	112.62	80.89
2	98.00	70.45

system ($p < 0.0001$). Follow-up tests are required to distinguish which specific system levels lead to significant differences in performance.

There are three search systems, corresponding to documents at the first rank position with relevance level 0, 1 or 2. However, for batch system performance to be expressed using the P@1 metric, relevance needs to be folded into a binary scale, giving two ways of grouping relevance levels: folding level 1 and 2 documents together, as is commonly done in TREC; or, folding level 0 and 1 documents together. Differences between these system levels are examined using the Wilcoxon signed-rank test, a non-parametric test of the null hypothesis that the median values of two samples are the same. For both relevance groupings, the differences in search times are statistically significant ($p = 0.0002$ for 0 versus 1 and 2; $p < 0.0001$ for 0 and 1 versus 2) indicating that P@1 batch results transfer to the user task.

For DCG@1, multiple levels of relevance can be accounted for explicitly in the batch metric, so all three systems can be compared directly. User performance differs significantly between systems 0 and 2 ($p < 0.0001$), and between systems 1 and 2 ($p = 0.0046$). However, the difference between systems 0 and 1 is only weakly significant ($p = 0.0989$). These results indicate that there is a noticeable difference for the average time that users need to find a useful document using search systems with different DCG@1 levels. However, this effect is most noticeable when comparing non-relevant documents (system 0) and highly relevant documents (system 2). Marginally relevant documents (system 1) are also clearly differentiated from highly relevant documents, but are similar to non-relevant documents.

These differences between the three relevance levels strongly suggest that, for P@1, it is preferable to fold level 1 (marginally relevant) documents with level 0 (non-relevant) documents, since the difference between level 1 and 2 is much stronger than the difference between level 0 and 1.

4.2 Relevance Profiling Based on Split Agreement

To investigate the effect of relevance mismatch on the relative outcomes of batch and user experiments, we classify users based on their relevance preferences expressed during searching. While working through the search topics, users were able to view documents, and then either choose to save them (a relevance vote), or close them and continue searching (a non-relevance vote). We use these relevance decisions to classify users, using the split agreement approach outlined in Section 2. Recall that users are classified into three groups: TREC-like (their relevance profile matches the TREC judging scheme); generous (their threshold

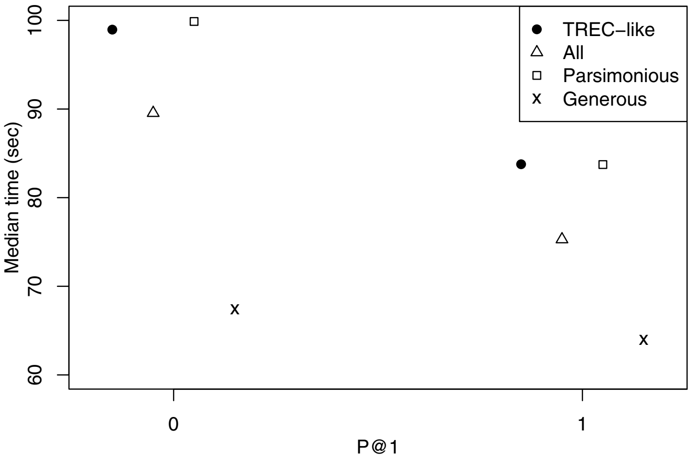

Fig. 1. Users categorised using the *split agreement* approach

for relevance is lower than that for TREC, so they are likely to accept level 0 documents as useful); and parsimonious (their threshold for relevance is higher than that for TREC, so they are unlikely to accept level 1 documents as useful).

If users can be successfully classified according to their relevance behaviour, then we would expect TREC-like users to show a larger difference in the time taken to find a useful document. That is, for users with relevance profiles that more closely match the criteria used in the batch experiment, the difference between retrieval systems as observed in the user task should be the most pronounced. Conversely, for users whose relevance profiles differ from the batch relevance judgements, the difference in retrieval systems should be less pronounced (generous users would be expected to be somewhat faster, no matter which system they are using; the opposite expectation holds for parsimonious users, who would be expected to be slower no matter which system they are using).

Note that here we are analysing relevance mismatch compared to the underlying batch experiment assumptions, based on the TREC relevance judgements: relevance is binary, with level 1 and level 2 documents grouped together into a single "relevant" category. That is, for the P@1 metric there are two possible outcomes, score of 0 (from level 0 documents), and a score of 1 (from level 1 or 2 documents).

Search times for different systems are shown in Figure 1. Across all users (represented by a triangle), task completion time falls when using a system with P@1 of 1, compared to 0. This difference is statistically significant, as shown in Table 2. Generous users are fast, whether they are using a system with a metric level of 0 or 1. As expected, the time taken to find a useful document is similar at both levels for this group, and the difference in batch metric does not lead to significantly different outcomes in the user task. Generous users are slowest when using the system where P@1 equals 0, and speed up when P@1 equals 1.

Table 2. Median time difference (in seconds) for a user to save a useful document for different levels of P@1 (p-values indicating the statistical significance of the time differences are shown in parentheses). Note that one user is in two classes.

User class	Median time difference (sec)		Number
All	14.26	(0.0002)	37
TREC-like	15.19	(0.0770)	8
Parsimonious	16.14	(0.0029)	11
Generous	3.43	(0.2209)	19

Users in the TREC-like class exhibit similar behaviour to the generous class. Both of these classes show a substantially larger difference in median time between the two systems than do generous users.

We note that, based on our post hoc grouping of users, the number of subjects in each class differs (for example, only 8 out of 37 users are in the TREC-like category, contributing to a weaker p-value despite the noticeable difference in median time). Nevertheless, it appears that relevance profiles can help to determine whether conclusions about batch P@1 values can be transferred to a user population: for generous users (who are satisfied with low-relevance documents as measured on the TREC scale, and don't differentiate strongly between any document levels), the differences are unlikely to hold. However, when the population consists of TREC-like or parsimonious users, the batch results are likely to be transferable.

5 Discussion and Conclusions

Batch evaluation is the dominant paradigm used to compare the performance of information retrieval systems. While a growing body of literature has suggested that there are mismatches between batch experiments based on widely used performance metrics such as MAP and actual search tasks are carried out by users, our results indicate the a simple performance metric such as P@1 can lead to search scenarios where the expected outcomes from batch experiments transfer directly to a precision-based user search task. This effect is statistically significant when relevance is treated as a binary criterion, as in the TREC framework. When multiple-level relevance judgements are available, DCG@1 is similarly effective at transferring expected batch experiment outcomes to a precision-oriented user search task. The difference in user performance is significant between the level 0 and 2, and level 1 and 2, relevance levels. However, it is only weakly significant between relevance levels 0 and 1. This suggests that the when multiple levels of relevance are folded into a binary scale, marginally relevant documents (level 1) should be grouped with non-relevant documents. This is in contrast with current standards used in IR evaluation, where marginally relevant documents are generally bundled with highly relevant documents.

The three system levels described above are intended to reflect possible scenarios of the DCG@1 and P@1 metrics. We note that, given the fixed distribution of relevance levels after rank position 1, our three defined *system levels* also correspond to particular values of other batch metrics. This holds for any metric that is only dependent on the relevance values of items within the top 10 positions of the ranked list (for example P@N or DCG@N for $N \leq 10$). However, for these metrics, the system levels defined for this study represent only a small range of the possible values that the metrics can take on. Therefore, the above conclusions from focusing on N=1 metrics should not be extended to the $N > 1$ alternatives directly. Moreover, metrics such as MAP, which include a recall component, will differ from topic to topic, since each topic considered will have a varying number of total relevant documents available. The conclusions from our experiments therefore do not transfer directly to such metrics.

We also investigated relevance mismatch, using split agreement to classify users into different relevance groups. Our analysis demonstrated that the transferability of batch experiment conclusions can differ between user classes; in particular, generous users, who have low thresholds for considering a document to be relevant, do not reflect the batch conclusions obtained form the P@1 metric.

The relevance profile analysis used the entire data obtained from the searching task; however, to be useful from a practical point of view, relevance matching should allow us to infer whether batch results are likely to successfully transfer to users, *without* requiring a full-scale user-study. In future work, we intend to investigate suitable approaches for estimating user relevance profiles with a minimum of effort. Naturally, there are many other possible causes of mismatch between batch experiments and user-based evaluations, including different levels of knowledge about the topics being searched on, age differences, cultural differences, and gender differences. We plan to incorporate these into the user classification approaches in future work.

References

1. Al-Maskari, A., Sanderson, M., Clough, P.: The relationship between IR effectiveness measures and user satisfaction. In: SIGIR, Amsterdam, Netherlands, pp. 773–774 (2007)
2. Allan, J., Carterette, B., Lewis, J.: When will information retrieval be "good enough"? In: SIGIR, Salvador, Brazil, pp. 433–440 (2005)
3. Borlund, P.: Experimental components for the evaluation of interactive information retrieval systems. Journal of Documentation 56(1), 71–90 (2000)
4. Buckley, C., Voorhees, E.M.: Evaluating Evaluation Measure Stability. In: SIGIR, Athens, Greece, pp. 33–40 (2000)
5. Buckley, C., Voorhees, E.M.: Retrieval system evaluation. In: Voorhees, E.M., Harman, D.K. (eds.) TREC: experiment and evaluation in information retrieval. MIT Press, Cambridge (2005)
6. Clarke, C., Craswell, N., Soboroff, I.: Overview of the TREC 2004 terabyte track. In: TREC 2004, Gaithersburg, MD (2005)
7. Clarke, C., Scholer, F., Soboroff, I.: The TREC 2005 terabyte track. In: TREC 2005. National Institute of Standards and Technology, Gaithersburg (2006)
8. Hersh, W., Turpin, A., Price, S., Chan, B., Kraemer, D., Sacherek, L., Olson, D.: Do batch and user evaluations give the same results? In: SIGIR, Athens, Greece, pp. 17–24 (2000)

9. Huffman, S.B., Hochster, M.: How well does result relevance predict session satisfaction? In: SIGIR, Amsterdam, Netherlands, pp. 567–574 (2007)
10. Ingwersen, P., Järvelin, K.: The Turn: Integration of Information Seeking and Retrieval in Context. Kluwer Academic Publishers, Dordrecht (2005)
11. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Information Systems 20(4), 422–446 (2002)
12. Kelly, D., Fu, X., Shah, C.: Effects of rank and precision of search results on users' evaluations of system performance. Technical Report TR-2007-02, University of North Carolina (2007)
13. Rose, D.E., Levinson, D.: Understanding user goals in web search. In: WWW 2004, pp. 13–19. New York (2004)
14. Scholer, F., Turpin, A., Wu, M.: Measuring user relevance criteria. In: The Second International Workshop on Evaluating Information Access (EVIA 2008), Tokyo, Japan, pp. 47–56 (2008)
15. Sheskin, D.: Handbook of parametric and nonparametric statistical proceedures. CRC Press, Boca Raton (1997)
16. Sormunen, E.: Liberal relevance criteria of TREC – counting on negligible documents? In: SIGIR, Tampere, Finland, pp. 324–330 (2002)
17. Spink, A., Jansen, B.J., Wolfram, D., Saracevic, T.: From e-sex to e-commerce: Web search changes. IEEE Computer 35(3), 107–109 (2002)
18. Turpin, A., Hersh, W.: Why batch and user evaluations do not give the same results. In: SIGIR, New Orleans, LA, pp. 225–231 (2001)
19. Turpin, A., Scholer, F.: User performance versus precision measures for simple web search tasks. In: SIGIR, Seattle, WA, pp. 11–18 (2006)
20. Turpin, A., Tsegay, Y., Hawking, D., Williams, H.E.: Fast generation of result snippets in web search. In: SIGIR, Amsterdam, Netherlands, pp. 127–134 (2007)
21. Vakkari, P., Sormunen, E.: The influence of relevance levels on the effectiveness of interactive information retrieval. Journal of the American Society for Information Science and Technology 55(11), 963–969 (2004)
22. Voorhees, E.M.: Variations in relevance judgements and the measurement of retrieval effectiveness. Information Processing and Management 36(5), 697–716 (2000)
23. Voorhees, E.M., Harman, D.K.: TREC: experiment and evaluation in information retrieval. MIT Press, Cambridge (2005)