User performance versus precision measures for simple search tasks

Turpin, Andrew; Scholer, Falk

Turpin, & Scholer, F. (2006). User performance versus precision measures for simple search tasks. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retreival, 11–18. https://doi.org/10.1145/1148170.1148176
Document Version: Published Version

Published Version: https://doi.org/10.1145/1148170.1148176

User Performance versus Precision Measures for Simple Search Tasks

Andrew Turpin[*] Falk Scholer

School of Computer Science & IT
RMIT University, GPO Box 2476V
Melbourne, Australia, 3001.
{aht,fscholer}@cs.rmit.edu.au

ABSTRACT

Several recent studies have demonstrated that the type of improvements in information retrieval system effectiveness reported in forums such as SIGIR and TREC do not translate into a benefit for users. Two of the studies used an instance recall task, and a third used a question answering task, so perhaps it is unsurprising that the precision based measures of IR system effectiveness on one-shot query evaluation do not correlate with user performance on these tasks. In this study, we evaluate two different information retrieval tasks on TREC Web-track data: a precision-based user task, measured by the length of time that users need to find a single document that is relevant to a TREC topic; and, a simple recall-based task, represented by the total number of relevant documents that users can identify within five minutes. Users employ search engines with controlled mean average precision (MAP) of between 55% and 95%. Our results show that there is no significant relationship between system effectiveness measured by MAP and the precision-based task. A significant, but weak relationship is present for the precision at one document returned metric. A weak relationship is present between MAP and the simple recall-based task.

Categories and Subject Descriptors

H.4 [**Information Storage and Retrieval**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Performance, Design, Experimentation, Human Factors

Keywords

Search engines, information retrieval evaluation, user study

1. INTRODUCTION

The field of information retrieval has a well-established tradition of experimental evaluation, dating back to Cleverdon's "Cranfield" experiments [7], and continuing through the ongoing series of Text REtrieval Conferences (TREC) [1]. The general approach for evaluating *adhoc* retrieval, where a static collection is searched for documents that are relevant to previously unknown topic, requires: a *collection* of documents that is to be searched; a set of *queries* that represent user information needs and are run against the collection; and a set of *relevance judgements* that indicate, for each query, which documents satisfy the current information need and which do not. Evaluations are typically run as a *batch process*, where the retrieval system fetches a pre-specified number of answer documents for each query, with no user interaction. Performance is quantified using a variety of metrics derived from the number of relevant answers that have been found. Commonly reported measures include mean average precision (MAP), precision at 10 documents retrieved (P@10), and bpref (these metrics are defined in Section 2). Indeed, much IR research focuses on demonstrating improvements in these metrics.

However, recent studies have demonstrated that improvements in these metrics do not translate into a direct benefit for users. A study by Hersh et al. [13] shows that instance recall – where users try to identify different aspects of a question within a limited timeframe – does not improve with small increases in mean average precision of the underlying search system on the scale that is commonly reported in IR results. Allan et al. [1] confirm this result (using bpref), but also show that for larger, specific increases in bpref, users do benefit on the instance recall task. Turpin and Hersh [17] demonstrate a lack of improvement when users are engaged in a question answering task for a small number of questions.

A possible reason for the lack of correlation between underlying system effectiveness and user performance could be the nature of the search tasks that have been examined. Instance recall – as its name implies – is inherently recall-oriented [1, 13]. However, mean average precision, while including a recall component, evaluates systems predom-

[*] Supported in part by a Discovery Project Grant from the Australian Research Council.

[1] http://trec.nist.gov

SIGIR'06, August 6–10, 2006, Seattle, Washington, USA.
Copyright 2006 ACM 1-59593-369-7/06/0008 ...$5.00.

inantly using precision [6, 19]. Similar observations hold true for metrics such as P@10. A question answering task introduces additional complications into the retrieval process: users not only need to identify relevant documents, but are also required to extract specific factoids to answer detailed questions. This introduces additional cognitive load on the user that may be not be reflected in document level relevance judgements, or by extension MAP calculations.

We investigate user performance based on two much simpler tasks. The first is a precision-oriented task, requiring users to find one document that is relevant to the supplied query. Such a search task may be expected to more closely reflect the system effectiveness that metrics such as MAP and bpref are measuring: recent research has demonstrated that users focus on the top ranked answers when looking at a ranked list of search results [15]. Since users only need to find one document, the position of relevant documents in the answer list as captured in the MAP metric would intuitively seem important. The second task that we consider is a simple recall-based task, measured by the number of relevant documents that users can identify in a five minute time period. This is simpler than the previously investigated instance-recall tasks, in that we do not require users to find novel information with each document discovered; different documents, that could repeat previously discovered information, are possible answers. Although a lack of correlation between system effectiveness metrics and user performance for more complex search tasks has been demonstrated, we wish to investigate whether a relationship exists between the metrics and simpler search tasks, in particular, tasks that typify the millions of searches conducted on the Web each day.

We have modelled our user interface on the interfaces of popular web search engines such as Google, Yahoo or MSN. By presenting users with a web search engine interface, we hope to examine whether the batch precision measures predict user performance on a simple web search task.

In our experiments, users were required to find documents that were relevant to a query in a short amount of time. The effectiveness of ranked lists for users was controlled using MAP, so we could measure user performance as a function of effectiveness. In all of our experiments and analysis, however, we could find no correlation between system performance measured with MAP and user performance on the precision task, and only a negligible improvement in performance on the recall task when MAP is increased.

Related work, including previous user-studies and details of IR system effectiveness metrics, is reviewed in Section 2. We then provide details of our experimental setup including the search task, collections and topics, in Section 3. Results are presented in Sections 4 and 5. We discuss our results in Section 6, and present conclusions in the final Section.

2. RELATED WORK

Information retrieval has a strong history of experimental evaluation. Two main methodologies are *batch processing* evaluation, and *user-based* evaluation.

Batch Processing Retrieval Evaluation

In the *adhoc* or *batch processing* paradigm [7, 19], a set of queries is run against a static collection of documents. The task of a retrieval system is to identify those documents in the collection that are relevant to the query. For evaluation purposes, relevance judgements are used to determine which documents are correct answers, and which are not. That is, a human manually examines each answer that a retrieval system returns, and decides whether the document is relevant for the query.

To enable the comparison of different retrieval systems, various system effectiveness metrics have been proposed. Most metrics are based on two properties of the answer set: *precision*, which focuses on how early in the ranking relevant documents are returned, and is defined as the number of relevant and retrieved documents as a proportion of the total number of retrieved documents; and *recall*, which is concerned with the completeness of the answer set, and is defined as the number of relevant and retrieved documents as a proportion of the total number of relevant documents in the collection.

Mean average precision (MAP) is one of the most widely-used system metrics, and gives a single numerical figure to represent system effectiveness [6]. Average precision for a single query is calculated by taking the mean of the precision scores obtained after each relevant document is retrieved, with relevant documents that are not retrieved receiving a precision score of zero. MAP is then the mean of average precision scores over a set of queries. MAP is a popular metric, and has been shown to be stable both across query set size [4] and variations in relevance judgements [18].

MAP assumes that complete relevance information is available – that is, for each query, every document in the collection is examined and evaluated as being relevant or not relevant. As collection sizes continue to increase, obtaining complete relevance information becomes problematic. TREC uses a pooling approach, where only those documents that are returned as possible answers to a query by participating systems are manually evaluated – all other documents in the collection are assumed to be not relevant. While some relevant documents may thus remain unidentified, this approach has been demonstrated to work effectively for the comparison of different retrieval systems [20]. As collection sizes continue to increase, however, the proportion of unjudged documents also increases, introducing a risk that a significant number of relevant documents remains unidentified. To overcome this problem, Buckley and Voorhees have recently proposed the *bpref* measure [5]. This measure only uses information from judged documents, and is a function of how frequently relevant documents are retrieved before non relevant documents. For evaluations with complete relevance information, bpref and MAP are strongly correlated [5].

Precision can also be calculated at particular cutoff points in the ranked list of answers that is returned by a retrieval system. *Precision at 10 documents retrieved* (P@10) is obtained by calculating the precision of a result set considering only the first 10 items in the ranked list. P@10 is a popular measure because it reflects the default number of answers that are returned on a single result page by popular web search engines. *Precision at 1 document retrieved* (P@1) is calculated based only on the relevance of the first item in the answer list.

User-based Retrieval Evaluation

The user-based evaluation of retrieval systems is complementary to the batch processing approach; here the focus is generally on the end users of retrieval systems. The evaluation of users as they perform search tasks has been studied as part of the TREC conferences, first in a dedicated interactive track [12, 14], and later in interactive "sub-tracks" [8]. We focus here on studies that have investigated the relationship between batch processing metrics and usability as demonstrated by users engaged in different search tasks.

Hersh et al. investigated whether batch and user evaluations give the same results for an instance recall task [13]. For this type of search task, users are required to find and mark documents that contain as many different instances about a topic as possible. For example, for a topic "dangerous wildlife in Africa", users would need to identify documents that mention as many different types of dangerous African wildlife as they can. Users were presented with search results from two systems: a baseline system with a MAP of 0.275, and an improved search system with a MAP of 0.324. Despite the fact that the difference in MAP between the systems was statistically significant, there was no evidence of a corresponding difference in user performance.

Allan et al. [1] investigated performance for an instance recall task at the passage level; that is, users were required to identify particular passages in documents that are relevant to a topic. In contrast to the experiments of Hersh et al., answer lists were artificially created at different levels of system quality, as measured by the bpref measure. This enables the comparison of user performance across a large range of underlying system effectiveness levels – users were presented with lists that had a bpref in the range from 0.5 to 0.98. Allan et al. found that different levels of bpref can have a statistically significant effect on user performance, but only at certain ranges of bpref level. In particular, recall (normalized by the time taken to find the answers) is significant only between system bpref of 0.5 to 0.6 ("hard" topics), and between 0.9 and 0.98 ("easy" topics). For the intermediate ranges, there is no relationship between user performance and bpref.

The relationship between system effectiveness and user performance for a question answering task was considered by Turpin and Hersh [17]. In contrast to an instance recall task, here users were required to identify a number of factoid answers to a question, or to choose a correct response from two possible answers [12]. Two search systems, with MAP scores of 0.270 and 0.354, respectively, were evaluated; no significant improvement in user performance for the question answering task was observed.

While the Hersh et al. study showed no correlation between user instance recall and system MAP for low MAP values (typical of those reported in IR studies), the Allan et al. study showed that some correlation was present for higher MAP values. The Turpin and Hersh study showed no correlation between user's ability to answer questions with systems of differing MAP at low values, but it is unknown what happens on a question answering task, or a simple informational web search [3], when the effectiveness of the retrieval system increases to higher levels of MAP. We attempt to address this issue in this study.

3. METHODS

The lack of evidence for differences in effectiveness at the user level based on difference in system effectiveness as measured by metrics such as MAP and bpref is of concern, as ultimately it is end users that retrieval systems aim to satisfy. However, previous experiments have focused on user search tasks that may promote aspects of searcher behaviour that are different from what the system effectiveness metrics are aiming to capture. For example, instance recall is a recall-oriented search task. MAP, on the other hand, is mainly influenced by the precision of an answer list (in the full MAP calculation, recall is reflected to some degree since relevant items that are not retrieved contribute a precision of zero to the average; however, the precision of found relevant items generally dominates). This paper aims to assess whether a relationship exists between system effectiveness metrics and a simple, precision-oriented search task.

```
<num> Number:  456
<title> is the world going to end 2000

<desc> Description:
Identify individuals or groups
predicting the end of the world in
the year 2000.

<narr> Narrative:
References to the apocalypse are taken
as equivalent to "end of the world" and
are therefore relevant.  Documents that
give imprecise references to "those
who believe...", for example, are not
relevant.
```

Figure 1: A sample TREC topic. Only the `<desc>` and `<narr>` sections were presented to the user. The `<title>` field was used as a query to seed ranked lists.

The precision-based search task that is the focus of this study is designed to emulate a simple information finding task as might be conducted by a regular web user: to find a document that contains material relevant to an information need. We also consider a recall-based task, where users simply identify as many relevant documents as they can in a five minute period of time. In our experiments, users accessed systems operating at different levels of effectiveness.

Users

Users were recruited by placing flyers around RMIT University during enrollment time, and on student newsgroups. In total, 30 students were recruited; the user population included both postgraduate and undergraduate subjects, with a mean age of 23 years. All were naive to the intent and details of the experiment, but all had some searching experience, and English as their primary language. The project was conducted under the guidelines of the Human Ethics Committee of RMIT University, and written informed consent was provided by all participants.

Users were required to complete a pre-experiment questionnaire, to establish their level of experience in conducting on-line searches. Most of our subjects reported that they have a great deal of experience with web search engines; the median search frequency among our user population was at least one on-line search per day.

Every user underwent a training session of one hour to ensure they were familiar with the task and the user interface before beginning the experiments proper.

```
INPUT       ${r_1, \ldots, r_n}$ are the relevance indicators
            of a list of documents, $L$, where $r_i$ is 1 if
            document $i$ is relevant to the topic, else 0.
            $T$ is the target AP of $L$.
Step 1      Let $AP(L)$ compute the AP of $L$.
Step 2      Set $i \leftarrow 0$.
Step 3      While $|T - AP(L)| > 0.005$ and $i < 1000$ do
Step 3.1      If $(AP(L) < T)$ then
                Randomly choose $r_i = 0$ and $r_j = 1$
                such that $1 \leq i < j \leq n$.
              else
                Randomly choose $r_i = 0$ and $r_j = 1$
                such that $1 \leq j < i \leq n$.
Step 3.2      Swap documents $i$ and $j$ in $L$.
Step 3.3      If $|T - AP(L)|$ is the smallest seen so far
Step 3.3.1      Store the $L$ in $S$.
Step 3.4      Increment $i$.
OUTPUT      The best list seen: $S$.
```

Figure 2: Algorithm used to randomly permute ranked lists to achieve a target AP.

Collection and topics

Aiming to simulate a web search task, we used the TREC WT10g collection for our experiments, a 10 gigabyte subset of a 1997 snapshot of the Web [2]. This collection has 100 associated topic-finding queries, TREC topics 451–550, created as part of the TREC 9 and 10 web tracks. For our experiments, we selected a subset of 50 topics. As we aim to evaluate systems with a large range of MAP scores (discussed further below), we selected those 50 topics with the highest number of relevant documents available in the TREC relevance judgements.

Users were presented with the description and narrative section of each topic as an information need that needed to be satisfied. An example topic is shown in Figure 1. The title field of each topic was used to generate query biased summaries [16] which were presented to the users in response to their queries. Users did not see the title field directly.

System Effectiveness

Effectiveness was measured using average precision over the list of documents returned to the user. As our lists were all 100 documents long, average precision was calculated as

$$\text{AP} = \frac{1}{\sum_{i=1}^{100} r_i} \sum_{i=1}^{100} r_i \left(\frac{\sum_{j=1}^{i} r_j}{i} \right),$$

where r_i is 1 if document i is relevant to the topic, and 0 otherwise.

In order to control the AP of the ranked list of documents returned to users, we randomly constructed answer lists from all known relevant and irrelevant documents as judged by the TREC assessors. The lists were constructed by taking the list of documents as they appeared in the TREC *qrels* file, and applying the algorithm outlined in Figure 2 to achieve a given level of AP. Each system always returned a list with the same AP, hence the Mean Average Precision (MAP) of the system over a set of topics would be equal to the AP of the system.

Based on the findings of Allan et al., we used search systems with MAP levels of 55%, 65%, 75%, 85% and 95% for our experiments. We refer to each level of MAP as a search *system*. Two hundred result lists were randomly constructed using the algorithm for each combination of topic and system.

System	AP	Sessions
1	55%	232
2	65%	228
3	75%	226
4	85%	228
5	95%	227

Table 1: Number of sessions (user-topic pairs) recorded for each system.

Query biased summaries were pre-computed for each document in each list using the title field of the topic as the query terms. The summaries contained whole sentences from the original documents that contained some query words, with a bias towards the number of query words contained, and the proximity of the query terms within the sentence [16]. The document title and the query biased summary were returned to the user.

Experimental Design

The performance of users is evaluated based on a traditional web searching task: given an information need, the user needs to issue queries and then identify answer documents that contain information relevant to the information need. Specifically, users were required to identify as many relevant documents as possible within a five minute time limit.

Each of the 30 users completed 50 topics, 5 with each system, in a balanced, pseudo-random order that controlled for order effects such as learning and fatigue in both the topics and systems. Due to some unanticipated use of the browser by users, a small number of the sessions (user-topic) were removed when the log entries were clearly nonsensical (for example, the view of a document preceding the issuance of a query that returned that document). Due to a software error, three topics were excluded from the final analysis, leaving 47 topics in total. This left us with a mean number of users for each topic of 24.3 (±3.4) with each system-topic pair used on average 4.9 (±1) times. The total number of user-topic pairs (also referred to as a *session*) for each system are shown in Table 1.

The search process for an individual session proceeded as follows. The Firefox browser was used on an X-windows platform; all users had prior experience with this interface. First, the user was presented with an information need, consisting of the description and narrative fields of a TREC topic, as explained previously. A timer was started from the moment that the topic was displayed to the user. The user was then free to issue search queries to the retrieval system: any number of queries could be issued within the five minute time limit. In response to a query, a result list consisting of 100 ranked answers was displayed. For each query issued in the session, a random selection from the 200 possible lists for the appropriate topic-system pair was returned. The same list was returned for identical queries within a session.

Each answer in a result list consisted of the title of the document, together with a two sentence query-biased summary. The title was also a hyperlink that could be clicked

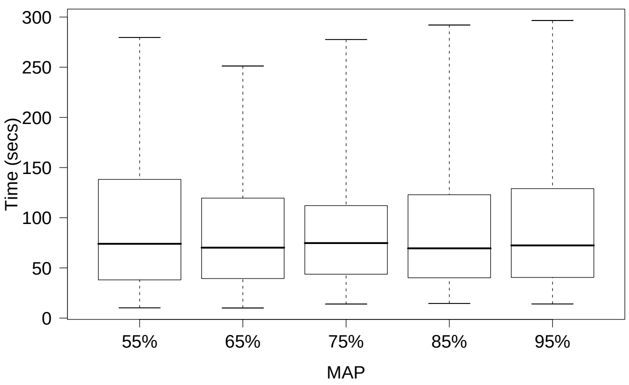

Figure 3: Time taken to find the first relevant document versus the mean average precision of the system used.

to view the actual document. The screen presented to the users resembled popular search engine screens without advertisements. The user was free to browse the answer list, and to follow the links to any documents that they wanted to examine in more detail. Following a link opened up the underlying document in a new window. After browsing the document, the user could choose to save the document if it was deemed relevant to the current information need by clicking a "Save" button, or could close the document window and return to the answer list.

At the end of five minutes the user interface returned to the start screen to prevent any further searching on the current topic by the user. Timestamps for all user interactions with the search system were recorded in a log file. The amount of time that a user took to correctly save their first relevant document measures their performance for our precision-based search task. The total number of relevant documents that were identified in five minutes measures user performance for the recall-based search task.

4. PRECISION-BASED SEARCH RESULTS

The relationship between the time that users took to find their first relevant document and other experimental factors are examined using a Multifactorial Analysis of Variance (ANOVA). The time taken to find a relevant document is the response against which the experimental treatments are analysed. Note that this response only includes sessions where at least one relevant document was saved, thus avoiding a ceiling effect introduced by the five minute time limit. The number of sessions where no relevant documents were found are examined separately below. Statistically significant effects occur within the set of users ($p \approx 0$), and within the set of topics ($p \approx 0$). However, there is no significant difference between the time taken to find the first relevant document using different systems ($p = 0.9291$). There were no significant interaction effects between users and systems, or between systems and topics.

Time to find a relevant answer

One of the simplest types of search task involves finding a single relevant document to satisfy an information need [3, 6]. Figure 3 shows a boxplot of the time taken by users to find the first relevant document versus system MAP, again excluding those times where no relevant document was found in a session. Boxplots in this paper show the median as the line in the box, with the box covering the 25% to 75% quartile of the data, and the whiskers extending to the extreme data values. As can be seen, the median time taken to find a relevant answer remains almost constant as the system MAP increases. This is consistent with the ANOVA result that there is no relationship between MAP and user performance as measured by the time taken to identify a relevant answer for an information need.

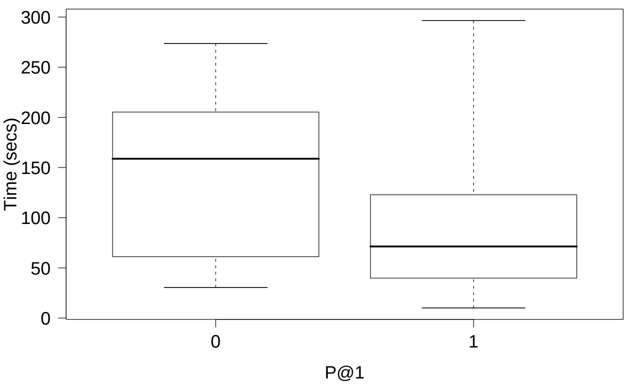

Figure 4: Time taken to find the first relevant document versus precision at 1 document retrieved.

System	MAP	No. Failures	Proportion
1	55%	116	21.4%
2	65%	122	22.6%
3	75%	100	18.5%
4	85%	99	18.3%
5	95%	104	19.2%
Total		541	100.0%

Table 2: Number of sessions where no relevant document was found in five minutes.

A similar lack of significant correlation was observed for the metrics P@2, P@3, P@4, and P@10. The median time taken to find a relevant document differs more strongly when the performance of the retrieval system is measured by P@1, as shown in Figure 4. This difference is statistically significant (Wilcoxon, $p = 0.011$). Note, however, that because only AP was controlled in any one session, it is possible for lists of differing P@1 to be seen by the user within one session. Whether the system had a P@1 equal to zero or one was determined from the list that contained the first relevant document saved. Thus the resulting difference for P@1 should be confirmed in a more controlled study.

Failure analysis

The preceding analysis excluded sessions where no relevant document was found within five minutes; this occurred in 541 of the 1147 total sessions (47.2%). These failed searches are shown, broken down by system, in Table 2. There is a small decrease in the number of failures between a MAP of 65% (System 2) and 75% (System 3), but none of the differences are statistically significant (chi-squared, $p = 0.19$).

Performing the same analysis for other metrics is more complicated as each session may contain several queries, and while the resulting list returned for each query has a controlled AP, it is not controlled for other metrics. For example, while it is possible to have an AP of 0.55 when all top ten documents are irrelevant, this is extremely unlikely to occur using the random list generation algorithm in Figure 2, and so a P@10 of zero was not observed in any of our lists. If we look at individual queries rather than sessions, however, we can say that of the 35 queries issued that returned a list with P@1 equal to zero, 24 (69%) of them did not result in a relevant document being saved. Of the 1848 lists returned to queries with P@1 of 1, 1225 (66%) of them failed to deliver a saved relevant document. The difference is small, but significant (chi-squared, $p < 0.001$).

User effect

A statistically significant effect is demonstrated among users. That is, averaged over all topics and all systems, some users take significantly longer to find relevant answer documents than others. This relationship is shown in Figure 6. The results suggest that there are large differences in the searching abilities of users who participated in our experiments; this is surprising, since our pre-experiment questionnaire suggested that most users had a relatively consistent level of prior online-search experience. It is possible that differences in reading comprehension, cognitive processing, or general language ability may have contributed to this effect. This observation could lead to interesting future research: investigating why users with similar levels of experience, and who use the same range of underlying search systems, nevertheless display great differences in performance; and examining how different groups of users can be best supported in their information finding tasks.

Topic effect

The ANOVA analysis of our experimental data indicates that there is a statistically significant topic effect. This is borne out by Figure 5, which shows the time required by users to find a relevant document for each topic, averaged across all five systems. The x-axis is sorted by the median time. Since each topic had an equal number of lists evaluated at each system level, this graph suggests that some topics are inherently more difficult than others. In particular, there is a noticeable jump in the median between the five rightmost topics and their left neighbors.

Motivated by this observation, we examined whether high MAP systems may be more effective on some topics than others, although the ANOVA had indicated that there was no significant interaction between topic and system. We created a set of “easy” topics, consisting of the result data from the five topics with the lowest median time required to find a relevant document (453—488), and a corresponding set of “hard” topics, using the data from the five topics with the highest median time (474—530). There are no observable or statistically significant correlations between system effectiveness as measured by MAP and time taken by users to complete a search task for the “easiest” (Pearson, $p = 0.8$) or “hardest” (Pearson, $p = 0.6$) five topics.

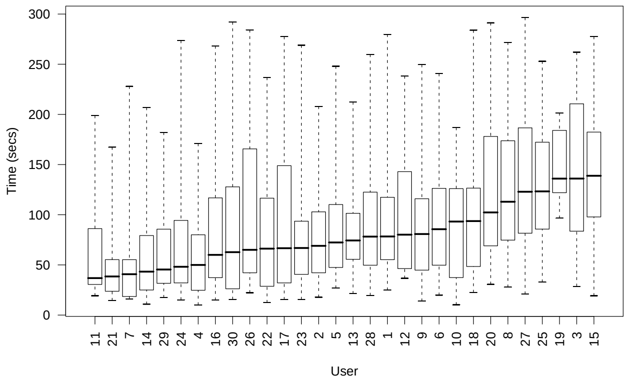

Figure 6: Time taken by different users to find the first relevant document, across all system levels and topics.

5. RECALL-BASED SEARCH RESULTS

The recall-based search task evaluated in our experiments is based on the number of relevant documents that users were able to identify within a five minute time period.

The distribution of the number of documents that users were able to identify is skewed heavily towards small numbers, and as such an ANOVA analysis is not appropriate. Instead we employed Tukey’s Honestly Significant Difference (HSD) test to analyze the statistical significance of effects.

Number of relevant documents found

Figure 7 shows a plot of the number of relevant documents that users were able to find in one five minute session using systems with different levels of MAP. From the plot it seems that once again there is no benefit in using systems with high MAP for this task. However, Tukey’s HSD test indicates that there is a significant difference in user performance between a MAP level of 55% and 75%, and between a MAP level of 65% and 75%. It seems that improving MAP up to 75% makes a statistically significant improvement in the systems ability to aid users in finding more than one relevant document in the five minute time limit. But in practice the effect is so small, 0.3 documents saved per session, on average, that it is unlikely to offer a real benefit.

User and topic effects

Similar to the precision-based search task, statistical analysis shows that there is a significant effect based on users for the recall-based task – that is, across systems and topics, some users are significantly better at identifying multiple relevant documents than other users are when carrying out the same task. In addition, statistically significant topic effects are observed, meaning that, across systems and users, some topics are inherently more difficult than others.

Answer ranks

Joachims et al. recently demonstrated that the position of a document in the answer list returned by a retrieval system can have a significant impact on whether a user is likely to view that document [15]. We therefore investigated the relationship between those documents that users viewed and

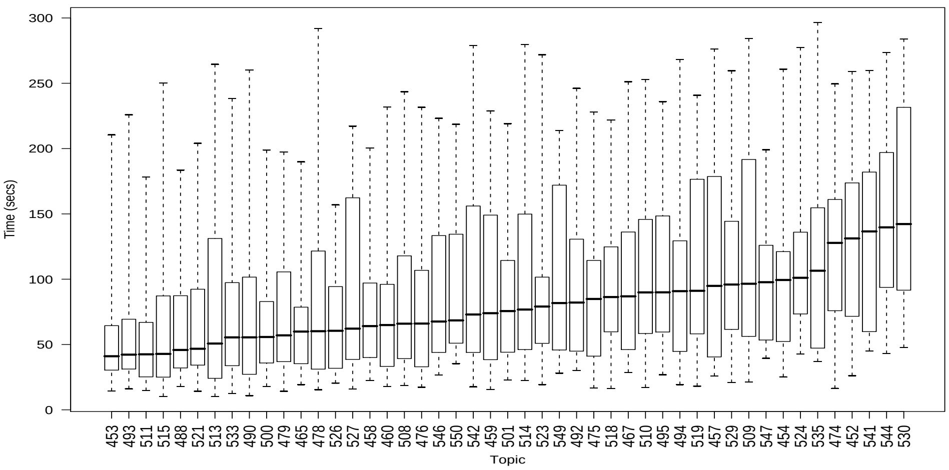

Figure 5: Time taken to find the first relevant document for different topics, across all system levels.

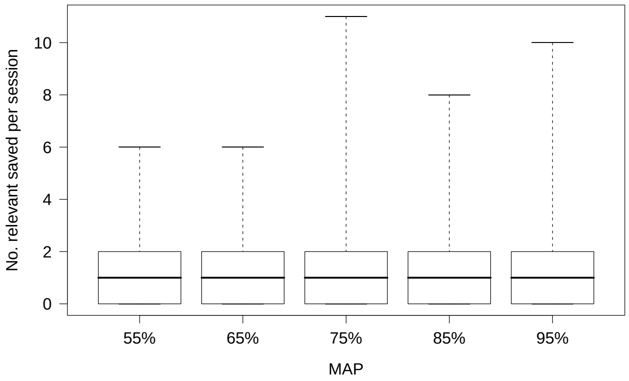

Figure 7: Number of relevant documents found by users within five minutes for systems with differing MAP.

saved, and the rank of those documents in our result lists. As found in the Joachims et al. study, user's interests are heavily focused towards the top of a list (ranks near 1), but, in our case, interest decreases almost linearly as ranks increase, unlike the dramatic decay in the number of clicks as the ranks increase reported by Joachims et al. A possible explanation for why our users more frequently clicked on documents further down the list than those reported in the Joachims et al. study could be the relatively high MAP of our answer lists. Joachims et al. used Google as their search engine, and assuming that it uses state-of-the-art retrieval techniques, MAP of greater than 50% is unlikely. Of course this conjecture can only be confirmed by performing relevance judgements on the lists returned to users in the Joachims et al. study.

6. DISCUSSION

Motivated by recent demonstrations that commonly reported IR system effectiveness metrics have little relationship to complex search tasks carried out by actual users, we used 30 users and 50 topics to investigate the relationship between such metrics on two relatively simple search tasks. The first was a precision-based task, where user performance was measured by the time taken to find the first relevant document; and the second a recall-based task, where user performance was measured by the number of relevant documents that could be identified in five minutes.

Our results demonstrate that there is little relationship between performance of systems as measured by MAP and the performance of users on these relatively straightforward search tasks. For the precision-based task, our analysis was not able to identify a relationship between the time that users take to find a relevant document and the MAP of underlying systems, or between the number of failures that users experience and the underlying system's MAP. This adds evidence to support the observation that "there is no single user application that directly motivates MAP" [6]. While MAP has been shown to consistently rank systems, it seems that the rankings achieved using this metric exhibit little relation to a ranking suggested by a user performing simple search tasks.

Users did find relevant documents more quickly when P@1 was one, as opposed to zero, but this result would require further investigation to confirm. Similarly, significantly less failures occurred with queries that returned a list with P@1 of one as opposed to zero.

There is a weak relationship between system performance measured by MAP and user performance on our recall-based task, suggesting that changes in MAP at moderate levels can lead to improved performance when users try to identify many relevant documents in a short period of time. However, the difference between systems was so small that it would be unlikely to make a difference in practice, and the difference was not noticeable above a MAP of 0.75.

One possible explanation for the lack of correlation between system precision and the ability of users to perform the tasks is the different environment in which the sets of relevance judgements were made for the two respective groups [11]. The MAP scores are calculated from TREC judgements, where a professional information assessor is required to assess the full text of the document relative to the topic and make a relevance judgement. Documents are presented to TREC assessors based on a document identifier (that is, in an arbitrary order) to remove ordering effects [10]. Our users, on the other hand, were presented with answer lists ordered by estimated relevance; the presentation order of the documents in therefore likely to have had an impact on the user relevance assessments [9].

Furthermore, the user judgements, are made in two stages: the first is a decision by the user whether to read the document or not, based on the summary presented; and the second is then to judge the document. If the first stage of triage is ignored, and only the second examined, our users judged 947 documents irrelevant (viewed, but did not save) that the TREC assessors judged relevant; and 80 documents relevant (viewed and saved) that the TREC assessors had deemed irrelevant. Of the 1867 (implicit) relevance judgements made by our users, only 840 (45%) agreed with the TREC assessors. This is consistent with Voorhees' study reporting agreement rates between assessors on TREC data [18].

Voorhees goes on to demonstrate that, even with two differing sets of judgements that overlap only by 32.8% – one from TREC assessors and one from a student group – the ranking of systems by the MAP metric are largely unaffected [18]. This implies that if we used judgements gathered from our users in the same environment as the test scenario, MAP levels might change, but the ranking of systems would remain similar to when TREC judgements were used. Hence there would still be no correlation between the performance of systems and the performance of users. It seems that the difference in judging circumstances may not be a suitable explanation for the lack of correlation between user performance and system performance measured using MAP.

For both tasks, strong effects were observed among the user factor – indicating that some users are better overall at searching than others – and among the topic factor – indicating that some search topics are more difficult than others. These observations suggest that further research is needed to establish what makes topics difficult for users (as opposed to difficult for systems ranked by their MAP score). While MAP and related metrics have been shown to be useful in comparing the relative performance of IR systems in batch mode, our findings cast further doubt on whether there is a direct relationship between these measures and actual user search tasks.

7. CONCLUSION

System performance as measured by a MAP of 0.55 and above does not correlate with user performance on simple information-finding web search tasks. P@1 shows promise as a metric that correlates system performance with user performance, but a study that controls P@1 explicitly is required to confirm this result. We hope that future research endeavors will consider the development of complementary IR evaluation metrics that reflect actual user performance.

8. ACKNOWLEDGMENTS

Thanks to Justin Zobel for helpful discussions, Pauline Chou for coding the user interface, Yohannes Tsegay for generating query biased summaries, and Steven Garcia and James Xiao for help with administering the user study.

9. REFERENCES

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

[2] P. Bailey, N. Craswell, and D. Hawking. Engineering a multi-purpose test collection for web retrieval experiments. *Information Processing and Management*, 39(6):853–871, 2003.

[3] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[4] C. Buckley and E. Voorhees. Evaluating evaluation measure stability. In *Proc. ACM SIGIR*, pages 33–40, Athens, Greece, 2000.

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

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

[7] 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).

[8] N. Craswell, D. Hawking, R. Wilkinson, and M. Wu. Overview of the TREC 2003 web track. In *The Twelfth Text REtrieval Conference (TREC 2003)*, pages 78–92, Gaithersburg, MD, 2003. NIST Special Publication 500-255.

[9] M. Elsenberg and C. Barry. Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *Journal of the American Society for Information Science and Technology*, 39:293–301, 1988.

[10] D. K. Harman. The TREC test collection. In E. M. Voorhees and D. K. Harman, editors, *TREC: experiment and evaluation in information retrieval*. MIT Press, 2005.

[11] S. P. Harter. Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, 47(1):37–49, 1996.

[12] W. Hersh and P. Over. TREC-9 interactive track report. In *The Ninth Text REtrieval Conference (TREC-9)*, pages 41–50, Gaithersburg, MD, 2000. NIST Special Publication 500-249.

[13] 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.

[14] W. R. Hersh. Trec 2002 interactive track report. In *The Eleventh Text REtrieval Conference (TREC 2002)*, Gaithersburg, MD, 2002. NIST Special Publication 500-251.

[15] 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.

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

[17] 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.

[18] E. M. Voorhees. Variations in relevance judgements and the measurement of retrieval effectiveness. *Information Processing and Management*, 36(5):697–716, 2000.

[19] E. M. Voorhees. The philosophy of information retrieval evaluation. In *CLEF*, volume 2406 of *Lecture Notes in Computer Science*, pages 355–370. Springer, 2001.

[20] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proc. ACM SIGIR*, pages 307–314, Melbourne, Australia, 1998.