The Good and the Bad System: Does the Test Collection Predict Users' Effectiveness?

Azzah Al-Maskari[1] Mark Sanderson[1] Paul Clough[1] Eija Airio[2]

lip05aaa@shef.ac.uk m.sanderson@shef.ac.uk p.d.clough@shef.ac.uk eija.airio@uta.fi

Dept. of Information Studies[1]
Sheffield, S1 4DP, UK
University of Sheffield

Dept. of Information Studies[2]
Tampere 33014, Finland
University of Tampere

ABSTRACT

Test collections are extensively used in the evaluation of information retrieval systems. Crucial to their use is the degree to which results from them predict user effectiveness. At first, past studies did not substantiate a relationship between system and user effectiveness; more recently, however, correlations have begun to emerge. The results of this paper strengthen and extend those findings. We introduce a novel methodology for investigating the relationship, which shows great success in establishing a significant correlation between system and user effectiveness. It is shown that users behave differently and discern differences between pairs of systems that have a very small absolute difference in test collection effectiveness. Our results strengthen the use of test collections in IR evaluation, confirming that users' effectiveness can be predicted successfully.

Categories and Subject Descriptors

H.3.0 [**Information Storage and Retrieval]:** General
General Terms: Measurement, Performance

Keywords: user study, Effectiveness measures, test collection

1. INTRODUCTION

Evaluation in Information Retrieval (IR) has a well established tradition in experimental design, with two main approaches dominating the field: system and user-oriented. The system-oriented approach, also known as batch-mode evaluation, typically involves a test collection: set of documents, set of information statements (topics), and relevance judgments for each topic. Systems are then evaluated in a lab-based setting by comparing documents returned for each topic with its set of relevance judgments. With careful construction, one can then presume that the performance of an IR system measured in the lab will reflect (or predict) its performance in practice (i.e. in an operational setting). The Cranfield experiments [4] were some of the first to develop and demonstrate the use of lab-based evaluation.

SIGIR' 08, July 20-24, 2008, Singapore.
Copyright 2008 ACM 978-1-60558-164-4/08/07…$5.00.

The effectiveness of IR systems is typically quantified using measures derived from the number of relevant documents returned by an IR system, or found by a user. Commonly used measures include Precision at rank 10 (P@10), Mean Average Precision (MAP), and binary preference (bpref). Much research in IR evaluation has focused on improving these measures, assuming that higher system effectiveness would help users to find more useful information, e.g. see [1, 5, 9, 10].

However, evaluation of this kind has been criticized by a number of researchers for the way it excludes users from the evaluation process [4]. This paved the way for a new direction in IR evaluation towards user-oriented approaches or interactive-mode [12]. Previous research [5, 9, 10] has demonstrated that improvements in system effectiveness do not correlate positively with users' effectiveness. Only in a more recent study by Allan et al. [1] correlations was substantiated. This is important because, as Järvelin & Ingwersen [7] assert, the real issue in designing IR systems is not whether precision or recall goes up by a statistically significant percentage, but rather whether it helps the user to solve the search task more efficiently and effectively. Interactive Information Retrieval (IIR) systems are those in which a user dynamically conducts searching tasks and reacts to system responses. Given the importance of test collections in IR evaluation, crucial to their use is the degree to which results from them predict user effectiveness.

This paper aims to build upon, and improve, the existing knowledge on whether or not test collections are applicable for predicting users' effectiveness. We describe a study in which 56 participants searched for 56 Text Retrieval Conference[1] (TREC) topics in two systems with different retrieval effectiveness. Users were required to save as many relevant documents as possible in response to their queries in a set time (a recall-based task). Results from this study show a significant relationship between system and user effectiveness: as the systems' effectiveness increases, users find more relevant documents, take less time, and report that they are more satisfied. This strengthens the use of test collections in IR evaluation, confirming that users' effectiveness can be predicted successfully.

The paper is structured as follows: Section 2 describes related literature in correlating user and system effectives, Section 3 details the experimental setup, including the search systems used by users, the search tasks and participants involved in this study,

[1] http://trec.nist.gov/ [site accessed: 23/01/08]

Section 4 discusses the results from our experiments, and Section 5 considers the implications of our findings on IR evaluation. We conclude the paper in Section 6.

2. LITERATURE REVIEW

A number of researchers in the IR field have questioned whether the results obtained from batch-mode evaluation can be generalized to predict the results of real interactive searching. Hersh, et al. [5] investigated whether batch and user evaluation gave the same results. In their study, 24 users attempted six instance-recall tasks using two systems with a MAP of 0.27 and 0.32 respectively. This study was performed in the context of the TREC-8 Interactive Track and although there was a significant difference in MAP between the two systems, there was no significant difference in the number of relevant documents saved by users searching with these systems.

In a subsequent study, Turpin & Hersh [10] performed further experiments based on a question-answering task to analyse the relationship between system effectiveness and user effectiveness. Again, 24 users were involved and required to identify a number of factoid answers to six questions from two systems with average MAP scores of 0.27 and 0.35. Despite the systems exhibiting quite different retrieval effectiveness, there was no significant improvement observed in user effectiveness (quantified by the number of questions answered).

Allan et. a1. [1] investigated system and user effectiveness for retrieval of document passages, by simulating lists of retrieval results at different levels of system quality (0.5 to 0.98 - as measured by bpref). They involved 33 users searching for 45 topics and results showed that differences in bpref could result in statistically significant differences in user effectiveness, as measured by the number of correct facts saved and time taken to find the answer. However, these results were only true at certain levels of bpref: the difference in number of facets saved was significant only between 0.50 and 0.80, and between 0.90 and 0.98; the difference in time taken to save the facets was significant only between 0.50 and 0.60, and between 0.9 and 0.98. For the intermediate ranges, there was no significant relationship between user effectiveness and bpref.

Turpin & Scholer [9] also created simulated lists of retrieval results with varying levels of MAP (0.55 to 0.95). Thirty users searched for 50 topics in both precision-based (time to find one relevant document) and recall-based (number of relevant documents found within five minutes) tasks. Results did not demonstrate a significant correlation between system effectiveness and user effectiveness in the precision task, and only a small improvement in the recall task.

Finally, Huffman & Hoshster [6] examined the relationship between session satisfaction and relevancy of the first three results to an information need. Seven participants assessed the relevancy of the first three results in response to 200 queries obtained from a Google log. Results from this study demonstrated that the higher the relevancy of the results, the stronger the participants' satisfaction.

While in the first two studies, [5, 10], two systems with different levels of effectiveness were used, they were limited to a small number of topics. This might explain why no correlation was detected. The following two studies, [1, 8], utilized systems with simulated retrieval results which had a substantial difference in systems effectiveness, however only in [1] did results prove a correlation between user effectiveness and system effectiveness. In this study we have followed a different experimental approach which has proven successful in establishing a link between system and user effectiveness. This helps corroborate the idea that test collections can indeed predict users' effectiveness.

3. METHODOLOGY

In devising the methodology, we wanted to present users with systems of varying effectiveness in a controlled way. In past experiments, users were presented with simulated lists of retrieval results with varying levels of system effectiveness. However, relying on a single system is not preferable as system effectiveness varies from one topic to another (a system with high effectiveness in one topic might be less effective for another). To provide a more controlled experiment, in our study we presumed that differentiating one system from another is better achieved on a topic-by-topic basis. Therefore, for this reason we utilized an experimental test bed, which allows access, through a single consistent interface, to three well known retrieval systems. We randomly selected 56 TREC topics and used the Title and Description fields from each topic as queries in the three systems. Each system gave different results and, using the TREC relevance judgments, we were able to compute Average Precision (AvP) for each system on each topic.

For each topic, two systems out of three were selected: the one with the highest AvP score and the one with the lowest AvP score, all AvP scores were averaged across the Title and Description. The system returning the highest number of relevant documents was categorized as the "good system"; the system returning the least number of relevant documents as the "bad system". Therefore, for each topic we had two systems: a good and a bad. The difference between the two on each topic provided a measure of system effectiveness against which to compare user effectiveness.

To obtain a measure of user effectiveness, interactive searches were performed for the 56 TREC topics. Users were required to search for topics using the good and bad system, but the effectiveness of the underlying system was unknown to them; thanks to the consistent interface of the test bed. With measures of system and user effectiveness for each topic, we were able to investigate the correlation between them. In this section, we outline the experimental set up which includes the systems, the task and the users.

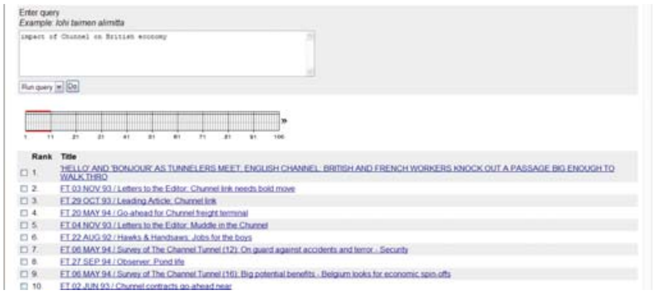
3.1 Retrieval System

The Query Performance Analyzer (QPA[2]), is a web-based application as illustrated in Figure 1. QPA provides a single interface to three different search engines: InQuery, Lemur and Terrier, which search the TREC-8 document collection. This test bed was used in our study. QPA records user interaction, but in addition we also observed users and recorded interactions using Camtasia Studio screen-capture software[3]. Each of the 56 topics was run in two systems, producing a total of 112 searches, which are distributed as follows: 43 in InQuery, 35 in Terrier and 34 in Lemur. The AvP for each system in response to these searches was 0.09, 0.12 and 0.16 respectively. The occurrence of these systems as being bad and good vary per topic: AvP for the bad

[2] http://www.info.uta.fi/julkaisut/pdf/qparn1.pdf [site accessed: 23/01/08]

[3] http://www.techsmith.com/camtasia.asp [site accessed: 23/01/08]

system is, on average, 0.05, and 0.20 for the good system. The quality of the bad and the good system varied highly across topics, from 0.000 to 0.417 within the bad system and from 0.001 to 0.843 within the good system, as shown in Figure 2.



impact of Channel on British economy

	Rank	Title
□	1.	'HELLO' AND 'BONJOUR' AS TUNNELERS MEET: ENGLISH CHANNEL: BRITISH AND FRENCH WORKERS KNOCK OUT A PASSAGE BIG ENOUGH TO WALK THRO
□	2.	FT 03 NOV 93 / Letters to the Editor: Chunnel link needs bold move
□	3.	FT 29 OCT 93 / Leading Article: Chunnel link
□	4.	FT 20 MAY 94 / Go-ahead for Chunnel freight terminal
□	5.	FT 04 NOV 93 / Letters to the Editor: Muddle in the Chunnel
□	6.	FT 22 AUG 92 / Hawks & Handsaws: Jobs for the boys
□	7.	FT 06 MAY 94 / Survey of The Channel Tunnel (12): On guard against accidents and terror - Security
□	8.	FT 27 SEP 94 / Observer: Pond life
□	9.	FT 06 MAY 94 / Survey of The Channel Tunnel (16): Big potential benefits - Belgium looks for economic spin offs
□	10.	FT 02 JUN 93 / Chunnel contracts go-ahead near

Figure 1: The QPA interface

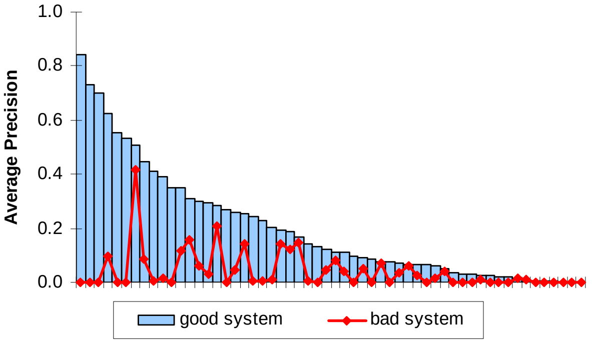


Figure 2: AvP score for each topic in the bad and the good system

3.2 Participants

In total, 61 students participated in our study, including both postgraduate and undergraduates in University of Sheffield. However, due to problems completing the task appropriately, five users were not considered in the analysis of results. Participants were required to complete a pre-experiment questionnaire to obtain demographic information and their level of experience in conducting on-line search. Most of our subjects reported that they had a great deal of experience with web search engines and that English was their first language or at least could be considered as their primary language. Prior to the actual experiment, every user underwent a training session to ensure they were familiar with the task and with the QPA system before starting the actual experiment.

3.3 Tasks

Participants were required to save as many relevant documents as possible for eight topics. Every eight users completed searches on the same set of topics: four in the bad system and four in the good system. Users were given 7 minutes for each topic and a Latin-Square arrangement was used to distribute the order of the topics among users (to reduce the effects of topic order on results). The amount of time given to users was determined from an earlier pilot study, conducted to make sure that users were given enough time to conduct the search effectively. Users were only presented with the first 200 results every time they conducted a search.

Users were presented with the description and narrative of TREC topics as an information need to be satisfied. Users did not see the title field to prevent them from using the title field as their query and in addition encouraging them to formulate their own queries. Users were free to issue as many queries as required for each topic within the 7 minutes. The narrative field served as guidance on the relevancy of the documents and users rated the relevancy of the documents using a ternary scheme: highly relevant[4], partially relevant[5] or not relevant.

After completing searching for each topic, participants filled out a questionnaire about their perception of the completed task. This requested information on the following: familiarity with the topic, easiness to complete the task, satisfaction with the results. All answers were reflected on a 4-point scale with 4 being the highest and 1 being the lowest.

Our hypothesis was that users would perform better by saving more relevant documents, spending less time, and hence be more satisfied with the good system than the bad system.

4. RESULTS AND ANALYSIS

Data in this study are reported using geometric mean to reduce the effects of outliers. This section is divided into five parts: 1) the effectiveness of users based on good and bad systems; 2) analysis of results at varying levels of difference; 3) the rank of relevant documents saved; 4) relationships between effectiveness measures; and 5) user's opinion about the search results. Analysis is based on grouping results in the following ways: in Sections 4.1, 4.2 and 4.3 analysis is based on combining results for the good and bad systems into separate sets; in Sections 4.4 and 4.5 analysis is based on combining results for both type of system into one set.

4.1 User Effectiveness

In this section we examine how user' effectiveness changed in response to using systems with different effectiveness: a good and a bad system. Table 1 illustrates a statistically significant difference (using a paired t-test) in system effectiveness for AvP between the good and the bad systems.

Table 1: Users' effectiveness and their perception in each system (N=56)

	bad system	good system	significance
AvP	0.05	0.20	0.00**
Time	1.83	1.47	0.02**
Rel docs	2.21	3.37	0.00**
Queries	4.09	3.27	0.00**
Easiness	2.10	2.48	0.00**
Satisfaction	2.04	2.34	0.00**
Rank_First	8.37	3.11	0.01**

**p<0.01

From Table 1, using the good system, users took less time to save the first relevant document, saved more relevant documents (that match with TREC relevance assessment) and issued fewer

[4] The document directly addresses the core issue of the topic

[5] The document only points to the topic, but it does a not discus the themes of the topic thoroughly

queries to complete the tasks. Consequently, users indicated that they found the task easier to complete using the good system, were more satisfied with results of the good system than the bad system and found the first relevant document at a higher rank than in the bad system.

4.2 Differences between Good and Bad Systems

The large difference in AvP scores, shown in Table 1, between the bad and the good system is likely to be responsible for the observed significant differences in user effectiveness. In this section we reduce the gap between the system effectiveness of the good and bad systems to determine the impact on user effectiveness. Reduction in the gap is achieved by subtracting AvP values of the good system from the bad system (absolute difference), per topic, and then sorting the results by the highest differences in AvP. Results are grouped into four equally sized sets; each containing the data points of 14 topics (Tables 2-5). The order of the attributes in these tables is consistently the same as in Table 1: *docs* is number of relevant documents saved, *Qs* is number of queries issued for each topic, *ease* is easiness to complete searching for topics, *sat* is users' satisfaction with the results, *RF* is ranking of the first relevant document for each topic, and *SS* is whether a statistically significant difference between bad and good systems exists. The idea behind this approach is to detect when significant differences in users' effectiveness between the bad and the good system disappear.

Table 2 has the highest difference between the good and bad system; while Table 5 has the lowest difference. We observe from these tables that as the difference between system effectiveness decreases, the difference in number of relevant documents saved also decreases until it gets insignificant (Table 5).

Table 2: Absolute Difference- set 1 (N=14)

	AvP	time	docs	Qs	ease	sat	RF
bad	0.05	1.91	3.16	4.08	2.27	2.21	2.21
good	0.47	1.23	5.00	2.64	2.51	2.35	2.42
SS	0.00^{**}	0.07	0.00^{**}	0.00^{**}	0.15	0.15	0.15

**p<0.01; *p<0.05

Table 3: Absolute Difference- set 2 (N=14)

	AvP	time	docs	Qs	ease	sat	RF
bad	0.04	2.04	2.18	4.14	2.27	2.20	8.58
good	0.19	1.48	4.08	3.52	2.67	2.70	3.31
SS	0.00^{**}	0.04*	0.01^{**}	0.13	0.15	0.02^{*}	0.03^{*}

Table 4: Absolute Difference- set 3 (N=14)

	AvP	time	docs	Qs	ease	sat	RF
bad	0.07	1.82	1.87	3.98	1.79	1.72	3.28
good	0.12	1.67	2.49	3.43	2.32	2.31	2.98
SS	0.00^{**}	0.33	0.05^{*}	0.07	0.01^{**}	0.00^{**}	0.36

Table 5: Absolute Difference- set 4 (N=14)

	AvP	time	docs	Qs	ease	sat	RF
bad	0.02	1.56	1.61	4.17	2.09	2.04	6.16
good	0.03	1.52	1.98	3.46	2.43	2.29	4.13
SS	0.00^{**}	0.45	0.11	0.05^{*}	0.04^{*}	0.14	0.02^{*}

Tables 6-9 show the relative difference between the good and the bad system, where results are also sorted by the highest difference in AvP. There is a consistent significant difference in the number of relevant documents saved from both systems, except in Table 9 where the relative difference is 30%.

Table 6: Relative Difference- set 1 (N=14)

	AvP	time	docs	Qs	ease	sat	RF
bad	0.00	2.64	1.50	4.20	2.02	2.03	16.85
good	0.30	1.39	3.92	2.74	2.44	2.35	3.28
SS	0.00^{**}	0.04^{*}	0.00^{**}	0.00^{**}	0.02^{*}	0.07	0.06

Table 7: Relative Difference- set 2 (N=14)

	AvP	time	docs	Qs	ease	sat	RF
bad	0.02	1.81	2.27	4.25	2.16	2.02	6.39
good	0.20	1.22	3.32	3.50	2.71	2.66	3.47
SS	0.00^{**}	0.04^{*}	0.01^{**}	0.07	0.01^{**}	0.00^{**}	0.06

Table 8: Relative Difference- set 3 (N=14)

	AvP	time	docs	Qs	ease	sat	RF
bad	0.05	1.44	2.61	3.92	2.25	2.14	6.32
good	0.15	1.41	3.55	3.18	2.42	2.43	3.12
SS	0.00^{**}	0.46	0.02^{*}	0.04^{*}	0.28	0.15	0.06

Table 9: Relative Difference- set 4 (N=14)

	AvP	time	Docs	Qs	ease	sat	RF
bad	0.11	1.37	2.32	3.98	1.96	1.97	3.74
good	0.14	1.83	2.50	3.56	2.35	2.27	3.19
SS	0.01^{**}	0.07	0.31	0.12	0.00^{**}	0.07	0.06

In the studies of [5, 10] there is a small significant improvement between their systems: (0.27 and 0.32) and (0.27 and 0.35). The absolute difference in their system effectiveness is even larger than the absolute difference in Table 11, but in their studies no significant improvement in the number of relevant documents saved or number of questions answered was observed. However, using the relative difference, the difference in study [10] is exactly the same (30%) as in Table 9. Their results are consistent with the results shown in Table 9 which lack significant differences between the two systems in the number of relevant documents saved. From the above 12 tables, we observe that users always save more relevant documents in the good system than the bad system provided that the absolute difference between them is larger than (0.01), and the relative difference is larger than 30% (Tables 5 and 9). This indicates that variations in system effectiveness can predict users' behaviour.

From the results in Tables (2-9), we observe that among the users' attributes, the significant difference of relevant documents saved gradually looses its significance as the gap between the bad and the good system lessens. However, the significant difference for other attributes (e.g. time, queries, satisfaction, and easiness) does not follow the same trend. We notice, particularly in Tables 2 and 9, that surprisingly that satisfaction is not significant, even with substantial differences between the two systems. Probably the lack of steady and consistent changes in these attributes, along with the difference between bad and good system, suggest they are not good measures (or rather they are "noisy") and cannot be predicted easily and accurately.

In the previous analyses, each table contained 14 topics only and one might argue that this is not enough to ensure reliable and valid results (given that Buckley & Voorhees [2] asserted that a good experiment should consist of at least 25 topics). Nonetheless, if we combine the results of Tables 2 & 3 (Table 10), Tables 4 & 5 (Table 11), Tables 6 & 7 (Table 12), and Tables 8 & 9 (Table 13), we now create four tables, each consisting of 28 topics.

Table 10 : Absolute Difference- set 1 (N=28)

	AvP	time	docs	Qs	ease	sat	RF
bad	0.04	1.97	2.37	4.11	2.27	2.21	12.02
good	0.33	1.35	4.51	3.08	2.59	2.56	2.46
SS	0.00^{**}	0.01^{**}	0.00^{**}	0.00^{**}	0.03^{*}	0.01^{**}	0.02^{**}

Table 11: Absolute Difference- set 2 (N=28)

	AvP	time	docs	Qs	ease	sat	RF
bad	0.05	1.69	1.79	4.07	1.94	1.88	4.72
good	0.08	1.59	2.24	3.45	2.38	2.30	3.55
SS	0.00^{**}	0.34	0.02^{*}	0.01^{**}	0.00^{**}	0.00^{**}	0.03*

Table 12: Relative Difference- set 1 (N=28)

	AvP	time	docs	Qs	ease	sat	RF
bad	0.01	2.26	1.95	4.23	2.10	2.03	11.53
good	0.26	1.33	3.72	3.16	2.58	2.51	3.06
SS	0.00^{**}	0.00^{**}	0.00^{**}	0.00^{**}	0.00^{**}	0.00^{**}	0.02^{*}

Table 13 : Relative Difference- set 2 (N=28)

	AvP	time	docs	Qs	ease	sat	RF
bad	0.08	1.40	2.46	3.95	2.11	2.06	5.33
good	0.15	1.62	3.02	3.37	2.38	2.35	3.16
SS	0.00^{**}	0.16	0.02^{*}	0.02^{*}	0.04^{*}	0.04^{*}	0.03^{*}

In tables 1-13, we only considered documents the users judged relevant and at the same time match with TREC relevance assessments. This is to be consistent with previous studies [1, 4, 5, 9, 10]. However, considering all the documents the users saved, both that match and do not match with TREC assessments, gives a different conclusion in the relationship between system effectiveness and users effectiveness (as measured by the number of relevant documents saved). At a large set of topics (N=56), users saved significantly (p<.000) more relevant documents in the good system (5.31) than the bad system (3.96). However, the trend disappears as number of topics is reduced to N=14 (although there is a significant difference at N=28 in the absolute difference, yet no significance difference in the relative difference). Hence, these results indicate that the sum of all the documents the users saved does not correspond with system effectiveness as compared with the documents that only match with TREC relevance assessment.

4.3 Rank of Saved Relevant Documents

In this section, we investigate the relationship between relevant documents saved by the user and rank of those documents in the results list. Joachims et. al. [7] demonstrated that the position of a document in the answer set returned by a retrieval system can have a significant impact on whether or not a user is likely to view that document.

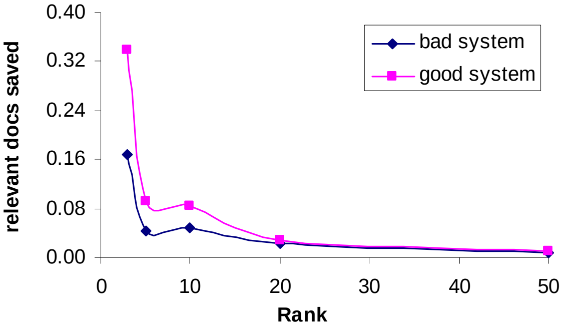


Figure 3: Proportion of relevant documents saved per rank

Figure 3 shows number of relevant documents saved for each of the following ranks: 3, 5, 10, 20 and 50. There is a substantial decrease in effectiveness from rank 3 to rank 50 in both systems (except at ranks 5 and 10 where users performed equally with lack of significant differences in the relevant documents saved). These findings support the results of [7] in that users' effectiveness at finding relevant documents decreases as they move down the rank. Comparing the effectiveness between the good and bad systems, there is a significant difference between them at ranks 3, 5 and 10 only; whilst they are the same from rank 20 downwards. It is surprising to see differences in number of relevant documents saved starts to disappear after the first 10 results. This signifies that after rank 10, users perform equally the same in systems with different effectiveness.

Figure 4 illustrates cumulative number of relevant documents saved for all the ranks: the highest number of relevant documents saved is at rank 3; the least saved is at rank 50. In the bad system, there is a lack of significance difference in the number of documents saved at rank 3 and rank 5. Notice that users' effectiveness in the bad system at rank 5 is equivalent to their effectiveness at rank 20 in the good system (whilst their effectiveness at rank 20 in the bad system is equivalent to their effectiveness at rank 50 in the good system).

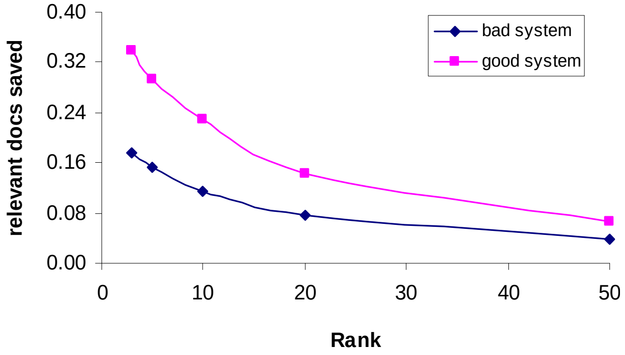


Figure 4: Cumulative no. of relevant documents saved

Table 14 illustrates that users search more effectively in the good system than the bad by retrieving and saving more relevant documents. But how true is this considering the absolute difference between relevant documents retrieved and saved in both systems? We argue that the smaller the difference between retrieved and saved documents, the better the effectiveness. According to the results in Figure 5, although users retrieved and saved more relevant documents in the good system, the absolute difference between retrieved and saved is smaller within the bad system than the good system. This observation highlights that the bad system is only bad at certain ranks: that is, not only within the first ten results but within the top three results.

Table 14: Retrieved to saved documents in both systems

	bad system		good system	
Rank	**retrieved**	**saved**	**retrieved**	**saved**
3	0.42	0.26	0.60	0.45
5	0.35	0.21	0.53	0.35
10	0.27	0.14	0.45	0.23
20	0.22	0.09	0.36	0.14
50	0.15	0.04	0.26	0.06

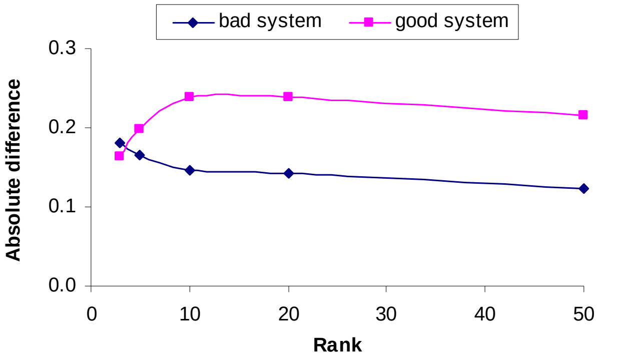


Figure 5: Absolute difference between retrieved and saved documents in both systems.

4.4 Effectiveness Measure Relationships

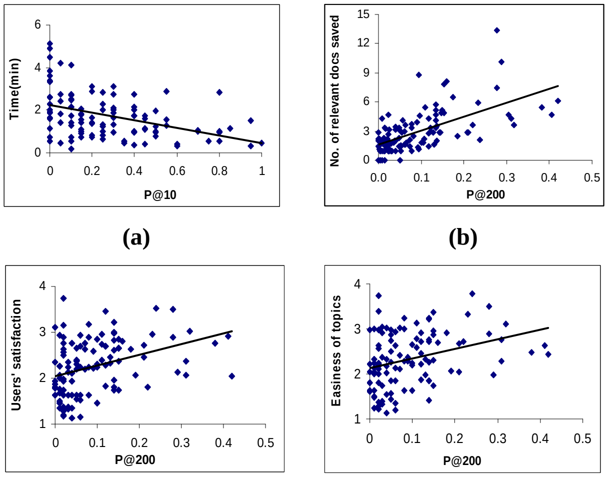
Results of the good and the bad systems were combined into one list to examine the correlation between system effectiveness measures: P@10, P@20, P@50, and P@200, and users' attributes (e.g. time, relevant documents saved, queries, satisfaction, and easiness), this is shown in Table 15. In general, all correlations reported here are low (the highest correlation is 0.62). We have highlighted the attributes with the highest correlation. It can be seen that the time to save the first relevant document correlates better with P@10. We speculate that the time correlates better with P@10 because users in general saved the first relevant document in the first 10 results: 86%, saved it at rank 6 (closest to P@10 than other measures) and 32% saved it at rank 1. However, it is surprising to see that P@200 correlates better with other users' attributes than AvP, although Buckley & Voorhees [2] had shown that MAP is more reliable and stable in comparison with others such as P@N. Perhaps this is because the Buckley & Voorhees [2] study is based on a batch-mode approach where no users were involved, whilst our study is conducted in an interactive mode. In a later publication, Buckley & Voorhees [3] stated that "there is no single user application that directly motivates MAP". Thus, MAP is a good measure in providing an overall view of system effectiveness, but P@200 is possibly a better indicator of user effectiveness and a more suitable measure for interactive IR evaluation.

From Table 15 we also observe that the number of relevant documents saved correlates better with all measures (P@10, P@20, P@50, and P@200) than any other attributes. The reason for the higher correlation with documents saved is because the scores for these measures are completely determined by the ranks of the relevant documents in the result. Moreover, these measures are derived in some way from precision and recall, which are both based on the proportion of retrieved documents that are relevant and relevant documents that are retrieved.

Table 15: Pearson's Correlation between measures of system and user effectiveness (N=112)

	P@10	**P@20**	**P@50**	**P@200**	**AvP**
Time	**-0.43****	-0.41**	-0.39**	-0.31**	-0.29**
Rel docs	0.52**	0.55**	0.58**	**0.62****	0.42**
Queries	-0.18*	-0.17*	-0.24**	**-0.29****	-0.27**
Satisfaction	0.29**	0.30*	0.33**	**0.36****	0.12
Easiness	0.26**	0.29*	0.30**	**0.33****	0.12

The low correlations in user attributes (Table 15) indicate that these attributes are hard to evaluate, even in a large study like this one. This is because users, in general, have different opinions about the relevancy of documents, even if they are provided with guidelines. Similarly, users' satisfaction is always likely to exhibit discrepancies amongst individuals. These variations make measuring users effectiveness challenging, and consequently hard to produce strong correlations. From these correlation, one can conclude that P@10 could be used if the intent of the study is to measure how fast it takes users to save the first relevant document (precision-oriented task) and P@200 could be used if the purpose is to measure the number of relevant document saved (recall-oriented task).



(a) **(b)**

(c) **(d)**

Figure 6(a-d): Users' effectiveness vs. system effectiveness (represents results from Table 8)

Figures 6(a-d) show scatter plots of effectiveness measures (i.e. P@10, P@200) plotted against the user attributes from Table 15. Correlation lines are overlaid on the plots. As can be seen, trends are present in the data, but the user measures, particularly time taken and user views on easiness of topic and overall satisfaction are all quite noisy, indicating the difficulty of measuring these attributes.

4.5 Users' Opinion about the Results

We have learned from above that P@200 correlates better with users' effectiveness than AvP. Therefore in this section, we decided to use P@200 as the measure of system effectiveness. We are investigating further at what levels of system effectiveness users are satisfied and not satisfied. Satisfaction is measured using a 4-point scale: 4 being the highest; 1 the lowest. Data points are divided into two sets: those with score bigger than 3 are grouped into "Satisfied"; those with score less than 3 are grouped into "Not Satisfied" (Table 16).

Table 16: Users' satisfaction per system effectiveness

	Satisfied (N=10)		Not satisfied (N=102)	
	P@200	**satisfaction**	**P@200**	**satisfaction**
Mean	0.14	3.29	0.09	2.16
Max	0.32	3.75	0.42	2.97
Min	0.00	3.00	0.00	1.13
correlation	-0.01		0.37*	

As can be seen in Table 16, on average users are satisfied if the P@200 is 0.14, and not satisfied if P@200 is 0.09. We have also obtained the correlation between users' satisfaction in the "Satisfied" and "Not Satisfied" sets. It is interesting to observe that the correlation within the Not Satisfied set (r=0.37) is stronger than the Satisfied set (r=-0.01). The higher correlation at systems with lower effectiveness possibly indicates that users agree on dissatisfaction when working with systems with low effectiveness; whilst their satisfaction varies more at higher effectiveness.

Table 17: Easiness to complete the task per system effectiveness

	Easy Topics (N=15)		Hard Topics (N=97)	
	P@200	**Easiness**	**P@200**	**Easiness**
Mean	0.13	3.28	0.09	2.17
Max	0.32	3.79	0.42	2.99
Min	0.01	3.01	0.00	1.13
correlation	0.28		0.33*	

The same procedure is followed in determining the easiness of completing the task (Table 17). Similar results to satisfaction are found: there is a stronger correlation between P@200 and easiness for the hard topics (r=0.33) than for the easy topics (0.28). This suggests that at systems with lower effectiveness, all users agree in the degree of difficulty with which relevant information is found, whilst in the better systems there is some discrepancy among individuals.

In the Turpin & Scholer study [9], they defined easiness according to the duration of the time required to save the first relevant document: the shorter the time, the easier the topic. They found no correlation between easiness and time to save the first relevant document. However, results from this study do not support their findings. Following the same approach in defining easiness, our results indicate a significant correlation between time and system effectiveness as measured by AvP, however low (r=-0.24, p=.01) and (r=-0.31, p<=0.01) when measured by P@200.

Furthermore, in the Allan et al. [1] study, easiness was defined according to system effectiveness (bpref): the higher the effectiveness, the easier the topic. They identified a relationship between easiness and system effectiveness, but only at certain levels of bpref (mainly between 0.50 and 0.60, and between 0.90 and 0.98). Applying their approach in our study would result in a correlation (r=0.33, p<0.000) between system effectiveness as measured by P@200. In either study, "easiness" has not been explicitly obtained from the users: in our study, we have asked the users to rate the easiness of each topic they have completed.

As can be seen from Table 18, easiness correlates higher with the relevant documents saved than with time to save the first relevant document.

Table 18: Correlation between users' effectiveness and their perception, (N=112)

	satisfaction	**easiness**
Rel docs	0.37**	0.34**
Time	-0.30**	-0.24**

Thus, we can suggest that number of relevant documents saved is the most suitable measure to represent users' satisfaction and topic easiness in an interactive study, especially when working with a recall-based task.

5. DISCUSSION

Previous studies considered the important problem of how improvements in system effectiveness reflect users' effectiveness, which means how test collection results predict users' effectiveness. Some of these studies used systems with simulated retrieval results, but struggled to identify significant relationships between system and user effectiveness. In a later study [1] some correlations started to appear. In this paper we have addressed the same issue in order to provide further evidence to this problem. However, we have followed a different approach which is distinctive from earlier studies. Our approach is also simulated in that system effectiveness is determined on a per-topic basis and not on system effectiveness across all topics (i.e. as a whole).

56 users conducted a recall-based task for 56 topics. Our users searched for these topics in two systems with different levels of effectiveness. Our findings showed that users saved more relevant documents, took less time to save the first relevant document and hence were more satisfied with systems of higher retrieval effectiveness than those with lower effectiveness. This indicates a relationship between system and user effectiveness. However, this relationship is true only if we consider the relevant documents the users saved that match with TREC assessments and ignore the ones that do not match with TREC assessment.

From this experiment we also observe that P@200 correlates better with users' effectiveness than average precision (AvP), although is recommended in system evaluation due to its stability. Currently we do not know what is causing the lack of correlation between AvP and users' effectiveness, but perhaps results suggest that AvP is a better measure in batch-mode evaluation; whilst P@200 would be better for interactive-mode. These results are supported by Turpin & Scholer [9] who find that the ranking achieved using MAP has little relation to ranking suggested by users performing a simple search task. We suggest that further research should be conducted to investigate this finding.

Although the number of relevant documents saved across systems correlates with system effectiveness, topic easiness and user satisfaction, it is still low (from 0.34 to 0.62). The system we used in this study does not provide snippets; only titles of documents. Therefore users first make a decision on whether to read the document or not, then if the title is promising users would click on the document to judge its relevancy. Nonetheless, if the title of the documents is not good enough yet the document itself is relevant, then the users would ignore it. Turpin & Hersh [11] indicated that showing only titles and not snippets did not have an effect to the users. However, we are not absolutely sure that this was the case and more investigation is required to reaffirm this.

A conclusion drawn from all the results shown here is that effectiveness measured on a test collection generally predicts a range of user effectiveness measures. However, the user measures are "noisy": 56 different users were employed in our study; users have different experience with search engines; each with different opinions on what a topic means and what documents are relevant. Such variation makes measuring users challenging, consequently, it is simply hard to find correlations. When differences between systems become slight, the variation in user behavior is too large to measure a user difference reliably. Many of the past papers that failed to find differences in user behavior across different search engines may simply have been encountering such noise.

However, it is important to note that these results may only apply to the tasks, topics and systems selected for this study. Precision-oriented tasks which require subjects to find one relevant document may have different results, and tasks that require subjects to work further down the ranked list would possibly have different implications. However, overall we believe this study supports the use of test collections in IR evaluation, confirming that users' effectiveness can be predicted successfully.

6. CONCLUSIONS

This experiment has added creditability to our hypothesis that users are able to complete retrieval tasks better when working with systems with higher effectiveness compared with systems of lower effectiveness. This leads us to conclude that users detect even small improvements in system effectiveness and this does impact on whether they complete a task successfully or not.

We have also learned from these experiments that a system with high effectiveness in batch-mode is superior to a less effective system, only at the first few rank positions in interactive-mode. This implies that after the first few results from a ranked list of search results, the usefulness of a system with high effectiveness is similar to one with low effectiveness.

Broadly speaking, this study has shown that better systems would result in better potency in users' effectiveness. Results from this study also show that P@200 is a good measure in interactive studies. Among the attributes considered in this study (number of relevant documents saved, time to save the first relevant documents and number of queries issued), we have found that number of relevant documents saved is a good measure to reflect system and user effectiveness, topic easiness and user satisfaction. We believe the results from this study will help researchers design more effective evaluation experiments.

7. ACKNOWLEDGMENTS

We would like to thank Ministry of Manpower, Oman, and the TrebleCLEF project (IST-FP7- 215231) for funding this study.

8. REFERENCES

[1] J. Allan, J., Carterette, B. & Lewis, J. When Will Information Retrieval Be "Good Enough"? User Effectiveness As a Function of Retrieval Accuracy. *ACM SIGIR*, pages 433-440, Salvador, Brazil.2005

[2] Buckley, C. & E. M. Voorhees. Evaluating Evaluation Measure Stability. *ACM SIGIR.* Athens, Greece .2000

[3] Buckley, C. & E. M. Voorhees. Retrieval system Evaluation. In E. M. VOORHEES & HARMAN D. K. (Eds.), *TREC: experiment and evaluation in information retrieval.* London, England, MIT Press. 2005

[4] S. P. Harter, & C. A. Hert. Evaluation of information retrieval systems: Approaches, issues, and methods. *Annual Review of Information Science and Technology (ARIST),* 32**,** 3-94. 1997

[5] 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.

[6] S. Huffman & M. Hochster. How Well does Result Relevance Predict Session Satisfaction?. *ACM SIGIR*.pages 567-573 Amsterdam, The Netherlands. 2007

[7] P. Ingwersen & K. Järvelin, *The turn: integration of information seeking and retrieval in context*, Springer. 2005

[8] 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.

[9] A. Turpin, & F. Scholer. User Performance versus Precision Measures for Simple Search Tasks. *SIGIR*, Pages 11-18. Seattle, Washington, USA. 2006

[10] A. Turpin, & W. Hersh. Why batch and user evaluations do not give the same results. *ACM SIGIR,* pages 225-231.New Orleans, Louisiana, United States. 2001

[11] A. Turpin, & W. Hersh. User interface effects in past batch versus user experiments. *ACM SIGIR,* pages 431-434. Tampere, Finland. 2002

[12] K. Yuri, & J. R. Moehr. Current Status of the Evaluation of Information Retrieval. *Journal of Medical Systems* 27(5): 409-424. 2003