Do User Preferences and Evaluation Measures Line Up?

Mark Sanderson, Monica Lestari Paramita, Paul Clough, Evangelos Kanoulas
Department of Information Studies, University of Sheffield
Regent Court, 211 Portobello St, Sheffield, S1 4DP, UK
+44 114 22 22648

m.sanderson@shef.ac.uk

ABSTRACT

This paper presents results comparing user preference for search engine rankings with measures of effectiveness computed from a test collection. It establishes that preferences and evaluation measures correlate: systems measured as better on a test collection are preferred by users. This correlation is established for both "conventional web retrieval" and for retrieval that emphasizes diverse results. The nDCG and ERR measures were found to correlate best with user preferences compared to a selection of other well known measures. Unlike previous studies in this area, this examination involved a large population of users, gathered through crowd sourcing, exposed to a wide range of retrieval systems, test collections and search tasks. Reasons for user preferences were also gathered and analyzed. The work revealed a number of new results, but also showed that there is much scope for future work refining effectiveness measures to better capture user preferences.

(*Note, this version replaces the copy of the paper found in the paper and CD proceedings of the ACM SIGIR 2010. In that version, tables 1 & 2 and 4-8 were at a late stage found to have an error, which this version corrects.*)

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]

General Terms

Measurement, Experimentation.

Keywords

Mechanical Turk, User Experiment, Evaluation Measures

1. INTRODUCTION

There is a long tradition of encouraging conducting, and researching evaluation of search systems in the IR community. A test collection and an evaluation measure are together used as a tool to make a *prediction* about the behavior of users on the IR systems being measured. If measurement using the collection reveals that system A is more effective than system B, it is assumed that users will *prefer* A over B in an operational setting. One of the striking aspects of almost all the early work in test collections is that the predictions about users implied from such measurements were rarely, if ever, validated. Given that test collections are used to simulate users, that so little validation took place is perhaps surprising.

In the last ten years a series of papers employing a range of methods conducted such validation. The papers produced contradictory results, some failing to find any link between test collection measures and user preferences, performance, or satisfaction; others finding links, but only when differences between IR systems were large.

Much of the past work involved a small number of topics, systems, and users; and/or introduced some form of artificial manipulation of search results as part of their experimental method. There was also a strong focus on test collections and not on the relative merits of different evaluation measures.

Therefore, it was decided to examine, on a larger scale, if test collections and their associated evaluation measures do in fact predict user preferences across multiple IR systems, examining different measures and topic types. The study involved 296 users, working with 30 topics, comparing user preferences across 19 runs submitted to a recent TREC evaluation. The research questions of the study were as follows

1. Does effectiveness measured on a test collection predict user preferences for one IR system over another?
2. If such a predictive power exists, does the strength of prediction vary across different search tasks and topic types?
3. If present, does the predictive power vary when different effectiveness measures are employed?
4. When choosing one system over another, what are the reasons given by users for their choice?

The rest of this paper starts with a literature review, followed by a description of the data sets and methods used in the study. Next, the results of experiments are described, the methods are reflected upon, conclusions are drawn, and future work is detailed.

2. PAST LITERATURE

The past work described here is grouped into two sections, based on the methods used to measure users. Contradictions between the results of the two groups are then discussed.

2.1 Measures rarely predict users

The power to predict user preferences using a test collection and evaluation measure was first examined in the work of Hersh et al [16] who used the 14 topics and qrels of TREC 6 and 7's interactive track to determine which of two retrieval systems was significantly better. They then conducted an experiment involving 24 searchers, retrieving over six topics of TREC-8: three topics on one system, three on the other. The researchers reported that there was no significant difference in the effectiveness of the searchers when using the different systems. This work was repeated on another test collection [26] drawing the same conclusion.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

Allan et al [4] created artificial document rankings from TREC data each with controlled levels of effectiveness. Users were shown selections of the generated rankings and asked to identify relevant information. Unlike the work described above, a correlation between user behavior and test collection based evaluation measures was found, but mainly when measured differences were large. Turpin & Scholer [27] repeated the artificial document ranking method, getting thirty users to examine fifty topics. No significant difference in the time users took to find the first relevant document was found. A small significant difference in the number of relevant documents identified was observed for large differences in the MAP of the artificial ranks.

Inspired by Hersh and Turpin' s method Al-Maskari et al [3] measured how well groups of users performed on two IR systems. Fifty six users searched from a selection of 56 topics. The researchers showed that test collection based measures were able to predict user behavior, and to some extent a user's level of satisfaction, however only when measured differences between the systems were large.

Although test collection based work is relatively recent, there is a longer tradition of correlating user outcomes with effectiveness measures calculated on actual searching systems. Tagliacozzo [23] showed that 18% of ~900 surveyed MEDLINE users appeared unsatisfied with search results despite retrieving a large number of relevant documents. As part of a larger study, Su [22] examined correlations between precision and user satisfaction; finding no significant link. Hersh et al [15] examined medical students' ability to answer clinical questions after using a medical literature search engine. No correlation between search effectiveness measures and the quality of the student's answers was found. Huuskonen et al [17] conducted a similar medical searching experiment reporting the same lack of correlation.

Smith and Kantor [21] engaged 36 users to each search 12 information gathering topics on two versions of a web search engine: one, the normal searching system and the other, a degraded version which displayed results starting from rank 300. Users weren't aware they were being shown the different versions. Although no actual effectiveness measures were taken, it is reasonable to assume that there was a significant difference in precision between the versions. However, there was no significant difference in user success in finding relevant items. Smith and Kantor reported that *users of the poorer system issued more queries*, which appeared to mitigate the smaller number of relevant documents retrieved in each search.

2.2 Measures predict user behavior

Measuring users through an analysis of query logs, Joachims [18] described an experiment showing users different sets of search results; as with previous work although there were measurable differences between the quantity and rank of relevant documents, Joachims saw little difference in users' click behavior. *Users given poorer search results still choose top ranked documents*. He proposed an alternative approach, which was to interleave the retrieval outputs of the two systems into a single ranking and observe if users tended to click on documents from one ranking more often than the other. The results showed users consistently chose documents from the better part of the interleaved ranking. This method of giving users (unknowingly) a choice and observing their preference was repeated [19] producing similar results. In this work, small, but measurable changes in document rankings were compared, and significant differences in user behavior were observed. Further analysis of query logs to model user click behavior was conducted by many researchers, e.g. [10].

Thomas et al [25] described another preference methodology where two sets of search results were presented side-by-side to users who were then asked which of the two they preferred. The method was used to compare the top 10 results of Google and the (presumably worse) Google results in ranks 21-30. They reported a clear preference for the top ranked results over the lower ranked.

2.3 Lessons drawn from past work

After reading the first set of research results, one might question the value of all test collection based research, as the only time users show any difference in behavior, success in their work, or preference for searching systems is when large differences in effectiveness between IR systems are measured. In direct contradiction to this, is the smaller body of work in the following section measuring clear preferences by users even for subtle differences in retrieval results. What might be the cause of this apparent contradiction?

Smith and Kantor's work appears to be the clearest in demonstrating that if it is important for users to locate relevant documents they can cope with the burden of a poorer search engine by re-formulating their query. In addition, Joachims' work appears to show that users will often make do with poorer results. The work in Section 2.1 could be failing to observe differences across users because these two traits simply make human searchers hard to measure.

As can be seen, there is only limited work using the preference based approach and to the best of our knowledge there is no work using this method to test the correlations between users and evaluations based on test collections. Further, none of the past work has addressed the more nuanced questions of whether certain evaluation measures or search tasks show better prediction of user behavior over others. Although there are a plethora of papers comparing different evaluation measures, almost without exception they report cross-measure correlations or use some form of stability statistic to imply which might be better. The only exception is Al-Maskari et al who examined correlations between user satisfaction and evaluation measures [2] finding that Cumulative Gain (CG) correlated better with user preferences than P(10), DCG and nDCG, but the experiment was based on a small sample of people.

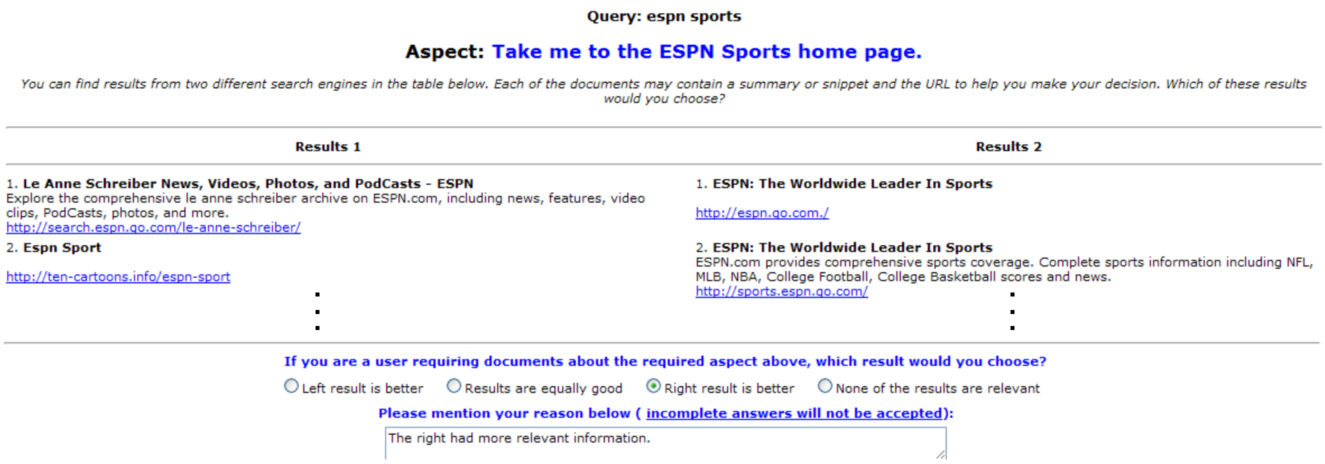
Because examination of different measures is almost unexplored, we addressed it here. With a growth of interest in search systems supporting diversity, there is as yet little research examining the predictive power of test collections in relation to diverse queries. Therefore, this paper conducted such a broad investigation into the predictive power of test collections and evaluation measures.

3. METHOD

The experiment required six components: a test collection with diverse topics and QRELS; multiple IR systems; a population of users; a method of measuring them; the selection of effectiveness measures; and a method of selecting which systems to show to users. These components are now described.

3.1 The test collection

The 50 million document Category B set of the ClueWeb09 collection was chosen as it was used in a diversity task for TREC's 2009 Web track. Given a short ill specified query, the



Query: espn sports

Aspect: Take me to the ESPN Sports home page.

You can find results from two different search engines in the table below. Each of the documents may contain a summary or snippet and the URL to help you make your decision. Which of these results would you choose?

Results 1	Results 2
1. **Le Anne Schreiber News, Videos, Photos, and PodCasts - ESPN** Explore the comprehensive le anne schreiber archive on ESPN.com, including news, features, video clips, PodCasts, photos, and more. http://search.espn.go.com/le-anne-schreiber/	1. **ESPN: The Worldwide Leader In Sports** http://espn.go.com./
2. **Espn Sport** http://ten-cartoons.info/espn-sport	2. **ESPN: The Worldwide Leader In Sports** ESPN.com provides comprehensive sports coverage. Complete sports information including NFL, MLB, NBA, College Football, College Basketball scores and news. http://sports.espn.go.com/
⋮	⋮

If you are a user requiring documents about the required aspect above, which result would you choose?

○ Left result is better ○ Results are equally good ● Right result is better ○ None of the results are relevant

Please mention your reason below (incomplete answers will not be accepted):

The right had more relevant information.

Figure 1 - Screen shown to MTurkers: containing query, subtopic, instructions, paired rankings, input buttons, and text box

goal of the diversity task was for participating groups to build IR systems that returned a ranked list of documents that collectively fulfilled the multiple information needs represented by the query. For the diversity track, each topic was structured as a set of subtopics, each related to a different user need [14]. The documents returned in the submitted runs were judged with respect to each subtopic. For each retrieved document, TREC assessors made a binary judgment as to whether or not the document satisfied the subtopic's information need.

Each one of the subtopics was categorized as being either *navigational* or *informational* (from Broder [9]). The query was also classified as either ambiguous or faceted, with ambiguous queries having multiple distinct interpretations while faceted queries had a single interpretation but with many aspects.

The structuring of subtopics judged in their own right into aggregated diverse topics, allowed (in this paper) both an experiment on diverse search and on non-diverse search: the first using the aggregated topics, the second treating the subtopics as a large set of ordinary topics.

3.2 IR systems

A source of different outputs was needed against which user preferences could be measured. Al Maskari et al in their experiments drew from a pool of three live searching systems, however, the researchers often found that the systems performed very differently from each other, which unsurprisingly resulted in large differences in user preference. In the design of the experiments here, it was judged desirable to have more explicit control over the differences between the systems being compared. Allan et al and others achieved this by artificially creating search results; we judged it preferable to use actual search output.

Arni et al [7] used the runs of an evaluation exercise as a source of search outputs to draw from to show users. From that pool of runs the researchers were able to select those runs that had similar effectiveness scores. For the category B ClueWeb09 collection, 19 diversity runs were submitted from ten research groups, these were the pool of search outputs used. Their use is detailed in 3.5.

3.3 Measuring user preference

To measure user preferences between different search results, the side-by-side method from Thomas et al [25] was chosen. For a particular topic, a pair of runs was selected from the pool and the top ten results (showing title, snippet and URL) were shown to users along with the topic title that generated the search and the subtopic description (referred to as an "aspect" in the interface) that expressed the information need behind the search request (example in Figure 1). The snippets were generated using a web service from the Bing search engine. Not all ClueWeb09 collection URLs still exist, which meant that 35% of results did not have a snippet. A post hoc analysis of data showed that missing snippets did not appear to influence user preferences.

Users were asked to indicate which of the two results they preferred. Using QREL data from the web track, effectiveness was measured on the two rankings and the agreement between users and the measures was assessed.

The aim of the diversity track was to promote searching systems that retrieved documents covering multiple interpretations of the same query, thereby ensuring that the search output was of value to the widest possible range of users. In a pilot experiment, an attempt was made to elicit user preferences for one IR system over another by asking individual users to indicate their preference for a ranking based on the ambiguous topic title alone. The expectation was that users would judge the value of search results relative to the multiple interpretations of a topic. However, it was found that the users were not able to do this reliably.

Therefore, in the experiments reported here, users were asked to focus on a particular subtopic and judge pairs of rankings in that context. They were asked to imagine they were searching for the subtopic using the query title text. The instructions were worded avoiding terms such as "diversity", so as not to bias choices. No other information about the experiment was given to the users. Users could indicate that the left or right result was better, both were equally good, or none of them were relevant (the ordering of paired systems was randomized). They were also asked to write a reason for their choice.

Different users were given the different subtopics of a topic and their preferences were aggregated to form a judgment on the diverse topic as a whole.

3.4 Population of users

The goal of the research work was to examine the preferences of a large number of users across many IR systems searching on a wide range of topics. It was decided to use the crowd sourcing system Mechanical Turk [5] to provide the large population.

Mechanical Turk users (*MTurkers*) were asked to judge a set of paired rankings for a set of subtopics. As it was assumed that there could be some disagreement amongst MTurkers, each pairing was seen on average by eight. A “trap question” was shown in an attempt to identify those who were not conducting the experiment in good faith. For every five comparisons shown to an MTurker one was a trap, which was built by pairing a run relevant to the required subtopic with a run for an entirely different topic. MTurkers who did not answer such pairings correctly had all of their answers rejected from the study (example in Figure 2). In total 342 MTurkers were used, 46 were rejected for failing a trap question (13%), which left 296 whose responses contributed to the results. We did not gather any demographic information from them. MTurkers were paid 8¢ for each block of five pairs they were shown. Many MTurkers worked on more than one block. The median time taken to complete the five pairs was just over 6 minutes. The total cost of the study including initial pilot studies was just under $60.

3.5 Selecting measures

The aim of the work was to examine how well evaluation measures predicted user preferences. Measures for both diversity and conventional IR were examined in this experiment.

3.5.1 Diversity measures

With the growth of interest in diversity, a number of evaluation measures were proposed. These measures include Cluster Recall (CR) used in ImageCLEFPhoto 2008 [7], Clarke et al’s α-nDCG [12], Agrawal et al’s intent aware Precision (IA-PC) [1], and Clarke et al.’s [13] novelty- and rank-biased precision (NRBP).

Cluster Recall (CR) is based on the subtopic recall (or S-Recall) proposed by Zhai et al. [29] to assess diversity. The CR at a cutoff rank *k*, CR(k), is defined as the percentage of subtopics covered by the first *k* documents in a ranked list. This is a pure diversity measure, i.e. it is not affected by the number of documents covering each cluster, or by their position in the ranked list. Further, it does not incorporate any notion of document novelty within a given subtopic.

Contrary to CR, both α-nDCG and NRBP consider both the number of relevant documents and their rank position over the subtopics of a query. For both measures, each document is assigned a gain value that is a function of the number of subtopics the document covers, and for each subtopic, the number of documents ranked above the given document that cover the same subtopic. The variable α is used to control how important diversity is in the measure. The α-nDCG metric is based on the traditional nDCG metric utilizing the aforementioned gain function, while the NRBP metric is based on Rank-Biased Precision (RBP). It is defined by replacing the traditional binary relevance of a document with the aforementioned gain. Thus, for a given subtopic, for α=0 the two metrics do not assess the novelty of the subsequent documents that cover this subtopic, while for α=1 they only consider relevant the first document that covers the given subtopic, ignoring all the subsequent ones.

Finally, intent aware Precision at rank *k* accounts for diversity and the number of relevant documents. Given a query, the precision value at cut-off *k* is computed for each subtopic separately (i.e. only the documents that cover the subtopic under consideration are considered relevant – called *aspect precision*) and the weighted average of the precision values is computed, with weights being the popularity of each subtopic. In the Web Track data the subtopics of a query were assumed to be equally popular.

3.5.2 Conventional evaluation measures

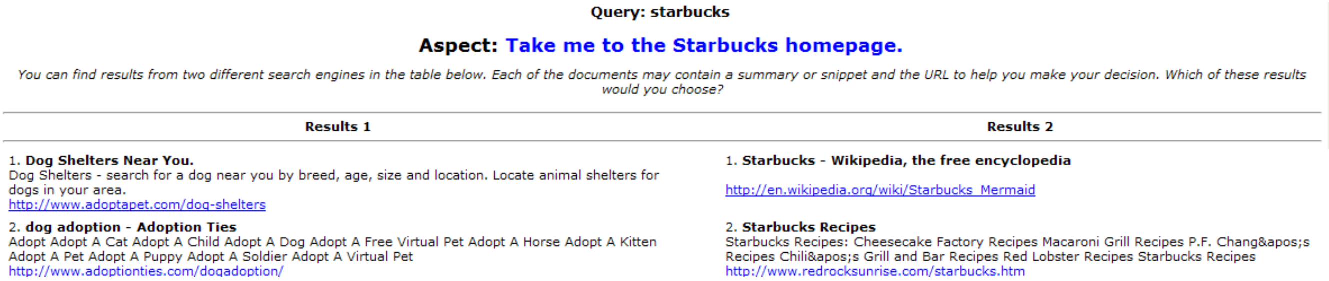
By considering each of the subtopics in the test collection as individual topics with their own QRELS, it was possible to examine differences across alternate conventional evaluation measures. Here nDCG, Mean Reciprocal Rank (MRR), Expected Reciprocal Rank (ERR) [11] and Precision measured at rank 10, P(10), were the measures chosen.

3.6 Selecting the pairs to show users

As seen in Section 2.1, existing research showed that differences in user performance could be measured on IR systems with large differences in search effectiveness. The challenge was in measuring user preference when searching on IR systems with far smaller differences. Therefore, the selection of run pairs to show the MTurkers focused on finding pairs that were similar to each other. There was a concern that using runs with low effectiveness could result in confusion when choosing between rankings. Therefore, topics where all runs had two or fewer relevant documents in the top ten were removed. This left thirty topics in the dataset.

A search was conducted across the remaining topics to locate pairs of runs that had the same number of relevant documents in the top ten, done to ensure that the rankings were similar. To enable diversity measures to be tested, runs were only paired when there was more than a minimum difference in subtopic coverage: ΔCR(10) and Δα-nDCG(10) ≥0.1. Runs submitted by the same research group were not paired together.

In total, 79 system pairs matching the search criteria were found. Each system pair was shown to, on average, eight MTurkers for each of a topic’s subtopics. The MTurker judgments for one of the 79 pairs were gathered as follows. Each system pair displayed the two retrieval results for a search based on the query text of a particular topic. The MTurkers were asked to indicate their preference for one of the paired systems in relation to a particular subtopic. Multiple MTurkers were shown the same



Query: starbucks

Aspect: Take me to the Starbucks homepage.

You can find results from two different search engines in the table below. Each of the documents may contain a summary or snippet and the URL to help you make your decision. Which of these results would you choose?

Results 1	Results 2
1. **Dog Shelters Near You.** Dog Shelters - search for a dog near you by breed, age, size and location. Locate animal shelters for dogs in your area. http://www.adoptapet.com/dog-shelters	1. **Starbucks - Wikipedia, the free encyclopedia** http://en.wikipedia.org/wiki/Starbucks_Mermaid
2. **dog adoption - Adoption Ties** Adopt Adopt A Cat Adopt A Child Adopt A Dog Adopt A Free Virtual Pet Adopt A Horse Adopt A Kitten Adopt A Pet Adopt A Puppy Adopt A Soldier Adopt A Virtual Pet http://www.adoptionties.com/dogadoption/	2. **Starbucks Recipes** Starbucks Recipes: Cheesecake Factory Recipes Macaroni Grill Recipes P.F. Chang&apos;s Recipes Chili&apos;s Grill and Bar Recipes Red Lobster Recipes Starbucks Recipes http://www.redrocksunrise.com/starbucks.htm

Figure 2 – Partial screen shot of a trap question shown to MTurkers

system/subtopic pair; although if an MTurker failed a trap question, their preference judgments were removed. MTurker preferences were treated as votes for one system or another normalized by the number of MTurkers who examined the system/subtopic pair.

This process was repeated for each of a topic's subtopics and the mean of the resulting normalized majority values was taken. The system that the majority of MTurkers preferred across the subtopics was selected as the best system for that topic. At the same time a diversity measure was calculated for the two system rankings, the best was the one with the highest effectiveness score. If there was a tie in scores, the pair was not considered. Across the 79 pairs, the number of times that MTurkers agreed/disagreed with the diversity measure was counted. If there was a tie, the MTurkers were judged to have said that the ranks from the systems were equal.

4. RESULTS

The predictive power of test collections/measures was examined on both the diverse web search topics (section 4.1) and their component subtopics (section 4.2).

4.1 User preferences in diversity search

The results of the initial experiment are shown in Table 1. As can be seen, there was a preference amongst users for systems that were measured to be more diverse. Assuming a null hypothesis that MTurkers saw no difference between the paired systems, and the level of agreement was simply due to chance; using a t-test[1] it was found that $p<0.05$; the null hypothesis was rejected and the level of agreement in Table 1 was found to be significant.

Users	**α-nDCG**		**Small Δ**		**Large Δ**	
Agree	50	64%	25	60%	25	69%
Rank equal	4	5%	2	5%	2	6%
Disagree	24	31%	15	36%	9	25%
	78		**42**		**36**	

Table 1 – Agreement differences in small and large Δα-nDCG

Next, the 78 pairs, without a tie in α-nDCG, were sorted by their difference; those pairs greater than the mean of the differences were placed in a large Δ bin; the others in a small Δ bin. The figures for user agreement are also shown in Table 1. Although the agreement appeared to grow as the size of difference between the two rankings increased, a significance test between large and small Δ showed $p>0.05$.

The range of different cluster evaluation measures described above were also examined, see Table 2. In percentage terms little difference was found between the measures, however, there were a large number of tied scores using IA-PC. It is notable that the measure CR provided as effective a prediction of user preference as the other measures. Cluster Recall is simply counting the percentage of topic interpretations that are covered in the ranking.

Users	**α-nDCG**		**CR**		**NRBP**		**IA-PC**	
Agree	50	64%	54	69%	51	65%	28	60%
Rnk eq	4	5%	4	5%	4	5%	1	2%
Disgree	24	31%	20	26%	24	30%	18	38%
	78		**78**		**79**		**47**	

Table 2 – MTurkers' agreements to the diversity measures

Given that we have observed similar degrees of correlation between different diversity measures and user preferences, we next investigated how these different measures correlated with each other. Given that these measures assess somewhat different aspects of system effectiveness, a strong correlation would indicate that better systems are good in all aspects of effectiveness assessed by these measures. A weak correlation will indicate that different users prefer different qualities of the ranked lists and the sets of users whose preferences agree with each individual measure do not fully overlap even though it so happens to be of similar size.

Kendall's τ	**CR@10**	**NRBP**	**IA-PC@10**
α-nDCG@10	0.7956	0.8523	0.8424
CR@10		0.7159	0.7219
NRBP			0.7010
AP-correl.	**CR@10**	**NRBP**	**IA-PC@10**
α-nDCG@10	0.6719	0.8736	0.7867
CR@10		0.6282	0.5492
NRBP			0.6839

Table 3 – Correlations between diversity measures

For each measure, we considered the mean values for all systems/runs submitted to the TREC track and over all 50 queries were calculated. For each two measures, we calculated Kendall's τ, and the AP-correlation [28], see Table 3. Kendall's τ is a function of the minimum number of pair wise adjacent interchanges needed to convert one ranking into the other. The AP-correlation is a similar metric, which however mostly accounts for the swaps towards the top of the system rankings, i.e. the disagreements over the top ranked systems. It can be seen that, there is a positive correlation among all measures, the strength of which however differs among different measures. In particular, the most correlated measures are α-nDCG and NRBP. IA-PC and α-nDCG are also well correlated, however, they mostly agree on the poorly performing systems as indicated by lower AP-correl. Further, there is a positive correlation between CR and α-nDCG; however it also concerns the bottom performing systems. Finally, CR and IA-PC correlate well regarding the bottom performing systems but they rank the top performing systems differently.

Therefore, the weak correlation among several of these measures indicates that indeed they assess different aspects of system performance. However, given the results in Table 2 it seems that all of these aspects are important for an average user.

4.2 User preferences in traditional search

If one treats each subtopic of the test collection as a distinct test collection topic, with its own QRELS, one can compare user preferences against traditional test collection measures. In total there were 250 subtopic/system pairs shown to MTurkers.

Three standard evaluation measures – nDCG, MRR, and P(10) and the newer measure ERR – were applied to the pairs and a prediction of which ranking users would prefer was made based on each measure. The standard measures were selected as exemplars of particular features in evaluation: P(10) is a simple count of the number of relevant documents in the top 10; MRR measures the rank of the highest relevant; nDCG combines both number of relevant documents and their rank. Both nDCG and ERR's ability to handle degrees of relevance was not exploited as the diversity track QRELS contained binary judgments only. For all measures, only the top 10 documents were examined.

[1] A 2 tailed, 2 sample unequal variance test was used throughout.

If the effectiveness measure for the two rankings were the same, user preferences were not examined. Therefore the number of pairs considered differed across the measures. For example, three pairs were measured as the same by nDCG; therefore, only 247 pairs were evaluated. The results of this analysis are shown in Table 4.

Users	**nDCG**		**MRR**		**P(10)**		**ERR**	
Agree	160	65%	159	67%	131	62%	164	66%
Rnk eql	21	9%	21	9%	18	9%	21	9%
Disgree	66	27%	57	24%	61	29%	62	25%
	247		**237**		**210**		**247**	

Table 4 – MTurkers's agreement with traditional measures

No significant difference in percentages between the measures was found. Focusing on nDCG, as in Section 4.1, the pairs were split into two bins: one for pairs with a large Δ and one for a small Δ. The split was defined by the mean difference between the 250 pairs. The figures for user agreement are shown in Table 5.

Users	**nDCG**		**Small Δ**		**Large Δ**	
Agree	160	65%	96	62%	64	70%
Rank equal	21	9%	16	10%	5	5%
Disagree	66	27%	43	28%	23	25%
	247		**155**		**92**	

Table 5 – Agreement differences in small and large ΔnDCG

Users appeared to agree more when there was a large difference in the evaluation measure than if there was a small one. However, as with the comparison in Table 1 no significant difference was found.

An alternate way to split the 250 pairs was on whether one of the two rankings contained no relevant documents in the top 10. Table 6 and Table 7 show the agreement figures based on this split. Contrasting the strength of user agreement between Table 6 and Table 7, for all columns MTurkers agreed more strongly when one pair of runs had a score=0. This was confirmed with a statistically significant difference being found between the figures in the first columns (nDCG) of the two tables.

Users	**nDCG**		**MRR**		**P(10)**		**ERR**	
Agree	88	72%	88	72%	88	72%	88	72%
Rnk eql	10	8%	10	8%	10	8%	10	8%
Disagree	24	20%	24	20%	24	20%	24	20%
	122		**122**		**122**		**122**	

Table 6 – Analysis of pairs with score=0 in one result

Users	**nDCG**		**MRR**		**P(10)**		**ERR**	
Agree	72	58%	71	62%	43	49%	76	61%
Rnk eql	11	9%	11	10%	8	9%	11	9%
Disagree	42	34%	33	29%	37	42%	38	30%
	125		**115**		**88**		**125**	

Table 7 – Analysis of pairs with score>0 in both results

An examination of mean Δ in nDCG between the pairs in Table 6 and Table 7 showed little difference, respectively 0.158 and 0.166. The best explanation for the difference was that it was due to the presence of a zero nDCG in one of the pairs. The results suggest a need for evaluation measures, e.g. GMAP [20] which penalize systems that fail to return any relevant documents for particular topics.

As would be expected for Table 6, the agreements were identical for all the measures examined. However, Table 7 showed key differences between the measures, particularly for P(10). Here the level of user agreement was almost random, this despite the measure scoring higher those ranking with more relevant documents. The highest levels of agreement were with MRR, but when the percentages were computed as a fraction of all pairs whose score > 0 including ties (resulting in a total of 128 pairs) nDCG and ERR appeared better, see Table 8. What appeared to be in little doubt was that P(10) was not well suited for assessing this sort of retrieval task.

Users	**nDCG**		**MRR**		**P(10)**		**ERR**	
Agree	72	56%	71	55%	43	34%	76	59%
Rnk eql	11	9%	11	9%	8	6%	11	9%
Disagree	42	33%	33	26%	37	29%	38	30%
Ties	3	2%	13	10%	40	31%	3	2%
	128		**128**		**128**		**128**	

Table 8 – figures from Table 7 with % based on all 128 pairs

The final analysis of this data was to examine different types of topic. Within the TREC Web collection, a small number of the subtopics were navigational, most were informational [9]. User agreement was measured split across these two topic types (see Table 9).

Users	**nDCG**		**Informational**		**Navigational**	
Agree	160	65%	146	64%	14	78%
Rank equal	21	9%	21	9%	0	0%
Disagree	66	27%	62	27%	4	22%
	247		**229**		**18**	

Table 9 – Analysis on different aspect types

For the small number of navigational topics, there was a strong agreement between users and the predictions made by the evaluation measures. No significance was found between the columns in this table. Given the small number of navigational topics, it would be valuable to repeat this experiment with an even balance in the number of navigational and informational topics.

4.3 MTurker comments on differences

In addition to indicating their preferences, MTurkers could also provide comments about their choices. In total, 96% of the judgments had associated comments that often indicated the reason(s) behind, or affecting, a decision. These often highlighted factors beyond the results simply having more relevant documents on a topic (informational) or a link to a required webpage (navigational). There were 11.6 words per comment, on average, and using an inductive approach to data analysis [24] comments in which the users made a specific preference (54% of those submitted, 1,307) were categorized. Fifteen classes were derived and in 88 cases, comments were assigned multiple categories, e.g. "*the left one has more useful results higher in the search*" was assigned the classes 'position' and 'number' indicating that the number of results and their position in the ranking would have likely influenced their preference judgment, see Table 10.

Although these are factors which researchers often highlight as affecting relevance [8], we see these mentioned unprompted by the MTurkers in this study, again highlighting the benefit of using MTurk to gather data for this kind of study, beyond implicit feedback strategies such as query logs.

5. REFLECTIONS ON METHOD

Here we discuss using MTurk as a source of user preferences; and using preference as a means of determining the impact of search on users.

5.1 Quality of the MTurk data

With an anonymous monetized crowd sourcing system, there is always a concern that data gathered will be overwhelmed with noise from spammers. However, evidence in our analysis such as time taken to complete tasks (median ~6 min.) of this set indicated that the majority of data was created in good faith. Indeed this gathering from hundreds of users of not only quantitative data, but also qualitative data gave this set a value that query/click logs do not have.

Nevertheless there are collections of data points in the set which we do not fully understand. Unexpected user responses to search results is not uncommon: Tagliacozzo [23], surveying ~900 MEDLINE users, described how nearly 18% declared dissatisfaction with search results despite earlier indicating that a large number of relevant documents were returned in the search they were asked to judge. Alonzo described how the development of MTurk experiments required multiple iterations to ensure that MTurkers were given clear instructions on what to do and to be certain that the responses from them are given in good faith [6].

The data set yielded a set of significant results on evaluation measures, but we view the method used here as a first step towards developing a more refined approach. Improvements will come not only from avoiding erroneous output from MTurkers, but also from building a more refined model of how users determine their preference for one searching system over another.

5.2 Does preference imply satisfaction?

The ultimate aim of this and past research was to understand if differences found between IR systems based on test collections were real: with a better system, would users be able to search more effectively, achieve their goals quicker, ultimately be more satisfied? Results from past work indicated that measuring such broader outcomes were challenging as users were adept at adapting either by searching more intensively or by managing to exploit relevant but poorer output.

The past work showed significant differences measured in a test collection did not necessarily show practical differences in user behavior. Was this simply because there was no practical difference to measure or was there simply a challenge in the way we measure users?

This and previous work showed that a preference based methodology can measure significant differences in users for relatively small differences in retrieval effectiveness. However, it is worth remembering that the side-by-side method is a simulation of search and the measurement of preference says nothing about whether users are more satisfied with their search or will achieve their information seeking tasks more effectively. One might wish to hypothesize that such a link exists, but testing that hypothesis is for future work.

6. CONCLUSIONS AND FUTURE WORK

The research questions posed at the start of the paper were answered through the gathering and examination of a large dataset of user preferences for a wide range of different IR systems tested over many queries and query types.

Clear evidence was found that effectiveness measured on a test collection predicted user preferences for one IR system over another. The strength of user prediction by test collection measures appeared to vary across different search tasks such as navigational or informational queries.

For diverse queries, little difference between diversity measures was found though the intent aware version of precision produced many ties between pairs. A conventional analysis of correlation between the measures was conducted confirming that they are similar. When comparing nDCG, MRR, ERR, and P(10) it was found that P(10) poorly modeled user preferences. However, user preferences between pairs of systems where one had failed to retrieve any relevant documents were notably stronger than when both rankings had at least one relevant document, which suggests a need for adjustments to these measures to allow for this user view.

Category	Example	#
On topic	“*All the results about secret garden*” / “*Contains work information in Michigan*.”	332
Specific links	“*the Wikipedia link for Hoboken will have most of the information I would be looking for*” / “*Right side has a link to the ESPN home page as asked for*.” / “*Right links to the desired information, the left does not*.”	265
Not classifiable	“*Each result would be helpful, but the left was easier*.” / “*I thought the left side was better*.”	181
Irrelevant	“*More non-relevant results in right column*.” / “*#5 on the left side is not for flame design at all*”	144
More relevant	“*more relevant*” / “*more relevant results on the right*”	132
Number	“*right has more map links*” / “*There are more results for finding houses or apartments in Hoboken*”	123
Position	“*Both lists include a link to reviews, but it's higher on the list on the left than on the right*.” / “*Top results more closely align to the request*.”	69
Range of results	“*The right column has a broader range of relevant topics*.” / “*seemed to include more of a variety of dino sites that would have what the person was looking for*”	66
Presentation	“*Results on the left are clearer and easier to read*.” / “*right results are more descriptive*”	36
Quality/ authority	“*right had a porn result, left is better*” / “*Left side has more relevant results, as well as listing more credible webpages*”	16
Spam/ adverts	“*Left seems to be legit. Right is more of junk and ads*” / “*Most of the right results are advertisements*” / “*less spammy*”/ “*Less commercial, more focused on the real subject results*.” / “*The right column just lists pet adoption classified ads*”	15
Duplication	“*right results has three repeated listings of Verizon wireless and broad band services” / “Almost all of the right results were just Wikipedia pages.*”	11
Availability	“*we can download the maps*” / “*more free data*”	6
Language	“*Left results have more relevancy, every link has something about pampered chef or related equipment in English*”	6
Dead links	“'*The right had dead links*” / “*The left had a few links which did not work but the majority did and returned results on appraisals*.”	5

Table 10 – Analysis of 1,307 MTurk comments

Finally, an examination and grouping of the written reasons that users provided for choosing one system ranking over another was outlined. Here it was shown that although relevance, rank, and information content of the documents was an important factor when users chose one system over another, a wide range of other reasons was also provided by users. Specific web sites were sought by some; avoidance of commercial sites or documents in other languages was desired by others. Such information from the MTurkers suggested that the test collections, relevance judgments and evaluation measures could be improved to provide a more effective model of user preferences than is currently available.

Such ideas are left to future work, where we will also continue to analyze our gathered data set as well as consider how to refine our collection methods to gather larger more informative sets in the future.

7. ACKNOWLEDGMENTS

We are grateful to William Webber and the Melbourne IR reading group for spotting an error in the earlier version of this paper.

This work was co-funded by the EU FP7 project ACCURAT contract FP7-ICT-248347 and Marie Curie Fellowship FP7-PEOPLE-2009-IIF-254562.

8. REFERENCES

[1] Agrawal, R., Gollapudi, S., Halverson, A., & Ieong, S. 2009. Diversifying search results. *ACM WSDM*, 5-14.

[2] Al-Maskari, A., Sanderson, M. & Clough, P., 2007. The relationship between IR effectiveness measures and user satisfaction. *ACM SIGIR*, 773-774.

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

[4] Allan, J., Carterette, B., & Lewis, J. 2005. When will information retrieval be “good enough”? *ACM SIGIR*, 433-440.

[5] Alonso, O., Rose, D. E., & Stewart, B. 2008. Crowdsourcing for relevance evaluation. *SIGIR Forum* 42, 2, 9-15.

[6] Alonso, O. & Mizzaro, S., 2009. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. *In Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*. 15–16.

[7] Arni, T., Tang, J., Sanderson, M., Clough, P. 2008. Creating a test collection to evaluate diversity in image retrieval. Workshop on Beyond Binary Relevance, *SIGIR* 2008.

[8] Barry, C. L. 1994. User-defined relevance criteria: an exploratory study. *J. Am. Soc. Inf. Sci.* 45, 3, 149-159.

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

[10] Chapelle, O. and Zhang, Y., 2009. A dynamic Bayesian network click model for web search ranking. *Proc. 18th WWW Conf*, 1-10

[11] Chapelle, O, Metzler, D., Zhang, Y., Grinspan, P., 2009. Expected reciprocal rank for graded relevance, ACM CIKM, 621-630

[12] Clarke, C., Kolla, M., Cormack, G., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I. 2008. Novelty & diversity in information retrieval evaluation. *ACM SIGIR*, 659-666.

[13] Clarke, C., Kolla, M., & Vechtomova, O. 2009. An Effectiveness Measure for Ambiguous and Underspecified Queries. *Advances in Information Retrieval Theory*, 188-199.

[14] Clarke, C., Craswell, N., and Soboroff, I. 2009. Preliminary Report on the TREC 2009 Web Track. *TREC 2009 Notebook*.

[15] Hersh, W.R. et al., 2002. Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions, *Am Med Inform Assoc*.

[16] Hersh, W., Turpin, A., Price, S., Chan, B., Kramer, D., Sacherek, L., & Olson, D. 2000. Do batch and user evaluations give the same results? *ACM SIGIR*, 17-24.

[17] Huuskonen, S. & Vakkari, P. 2008. Students' search process and outcome in Medline in writing an essay for a class on evidence-based medicine. *Journal of Documentation*, 64(2), 287-303.

[18] Joachims, T., 2002. Evaluating retrieval performance using click through data. *Workshop on Mathematical/Formal Methods in IR*, 12–15.

[19] Radlinski, F., Kurup, M., Joachims, T. 2008. How does click through data reflect retrieval quality? *ACM CIKM*, 43-52.

[20] Robertson, S. 2006. On GMAP: and other transformations, *ACM CIKM*, 78-83

[21] Smith, C.L. & Kantor, P.B., 2008. User adaptation: good results from poor systems. *ACM SIGIR*, 147-154.

[22] Su, L.T., 1992. Evaluation measures for interactive information retrieval. *IP&M*, 28(4), 503-516.

[23] Tagliacozzo, R., 1977. Estimating the satisfaction of information users. *Bulletin of the Medical Library Association*, 65(2), 243-249.

[24] Thomas, D.R. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*, 27(2), 237-246

[25] Thomas, P. & Hawking, D., 2006. Evaluation by comparing result sets in context. *ACM CIKM*, 94-101.

[26] Turpin, A.H., Hersh, W. 2001. Why batch and user evaluations do not give the same results. *ACM SIGIR*, 225-231.

[27] Turpin, A. & Scholer, F. 2006. User performance versus precision measures for simple search tasks. *ACM SIGIR*, 11-18.

[28] Yilmaz, E., Aslam, J., Robertson, S. 2008. A new rank correlation coefficient for information retrieval. *ACM SIGIR*, 587-594.

[29] Zhai, C.X., Cohen, W.W. & Lafferty, J., 2003. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. *ACM SIGIR*, 10-17.