A Comparison of User and System Query Performance Predictions

Claudia Hauff
University of Twente
Enschede, The Netherlands
c.hauff@ewi.utwente.nl

Diane Kelly
University of North Carolina
Chapel Hill, NC, United States
dianek@email.unc.edu

Leif Azzopardi
University of Glasgow
Glasgow, United Kingdom
leif@dcs.gla.ac.uk

ABSTRACT

Query performance prediction methods are usually applied to estimate the retrieval effectiveness of queries, where the evaluation is largely system sided. However, little work has been conducted to understand query performance prediction from the user's perspective. The question we consider is, whether the predictions of query performance that systems make are in line with the predictions that users make. To this aim, we compare the performance ratings users assign to queries with the performance scores estimated by a range of pre-retrieval and post-retrieval query performance predictors. Two studies are presented that explore the relationship between user ratings and system predictions on two levels: (i) the *topic level*, and, (ii) the *query suggestions level*. It is shown that when predicting the performance of query suggestions, user ratings were mostly uncorrelated with system predictions. At the topic level though, where a single query is judged for each information need, we observed moderate correlations between user ratings and a subset of system predictions. As query performance prediction methods are often based on intuitions of how users might rate queries, these findings suggest that such methods are not representative of how users actually rate query suggestions and topics. This motivates further research into understanding the rating process engaged by users, and developing models of query performance prediction in order to bridge the divide between systems and users.

Categories and Subject Descriptors

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

General Terms

Human Factors, Experimentation

Keywords

Query Performance Prediction, Query Suggestions, User Ratings

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

1. INTRODUCTION

An essential part of the Information Retrieval (IR) process is the formulation or generation of an effective query, i.e., one that retrieves relevant information. To determine whether a query is likely to be effective, there has been significant investment into the development of Query Performance Prediction (QPP) methods [8, 9, 14, 31, 36, 38, 39, 40]. These automated techniques aim to estimate the quality of the search results returned by a retrieval system in response to a query. The idea is that if an ineffective query can be detected *a priori*, measures can be taken by the system to improve the query and subsequently the search results. Conversely, if a query is predicted to be effective, its search results may be further improved by automatic query expansion.

The evaluation of the numerous QPP methods proposed in the past is largely based on their correlation with retrieval effectiveness measures such as average precision. However, it has been shown that such system oriented retrieval effectiveness measures can be poor indicators of actual user satisfaction [1, 16, 28, 30]. Perhaps the predictions made by QPP methods are more indicative of user ratings of query quality than retrieval effectiveness. Determining the extent of the relationship between the predictions of query quality[1] made by QPP methods and the quality ratings made by users is extremely important in the context of interactive and adaptive IR. For example, when offering query suggestions, the retrieval system needs to determine which suggestions the user is most likely to select. And, when users are presented with a set of query suggestions they are essentially being asked to assess a set of queries. While many researchers have examined the potential usefulness of term and query suggestions and techniques for automatically generating query suggestions in particular contexts [18, 26, 34, 35], little research has been conducted about how users make decisions about queries and suggestions, in contrast, to the system sided query performance prediction work.

Thus, an open question is whether the predictions about a query's quality by users are similar to the predictions made by the QPP methods employed by retrieval systems. To this aim, we compare the performance ratings users assign to queries with the performance scores estimated by a range of QPP methods. Two studies are presented that explore the relationship between user ratings and system predictions on two levels: (i) the *topic level*, and, (ii) the *query suggestions level*. At the topic level, the performance of a query

[1]In this paper, we use the phrases *query quality* and *query performance* interchangeably.

is estimated for each topic (i.e. information need), given a set of topics. This setup is the typical system sided QPP evaluation. At the query suggestions level, the performance of each query of a given set of queries is estimated for an information need. This is a setting that is more commonly experienced by users either implicitly (as they may have to think about posing several possible queries), or explicitly (as the retrieval system may recommend a set of suggestions). Thus, the query suggestions level can be considered more user oriented as it focuses on the problem of query selection. The studies performed as part of this research provide two distinct levels on which to consider these query performance prediction tasks. They allow us to better understand the relationship between users, retrieval systems and perceived/predicted query quality.

The paper is organized as follows: we first outline the two areas of related work, namely query performance prediction (Sec. 2.1) as well as term and query suggestions (Sec. 2.2). In Sec. 3 the user studies conducted and the QPP methods used in our experiments are described. Then, the results are presented and analyzed in Sec. 4, before Sec. 5 rounds off the paper with a discussion and the conclusions.

2. BACKGROUND

We first present an overview of QPP methods before detailing previous work on term and query suggestions. The first part is largely system oriented research, whereas the second part is more user oriented research. Although there is much overlap between the two areas, i.e., both systems and users making predictions about the quality of queries, little work has been conducted to consider both sides together.

2.1 Query Performance Prediction

Automatically predicting the performance (or retrieval effectiveness), of queries is a very active area of research [8, 9, 14, 36, 38, 39, 40]. This is because it is believed that accurately predicting a query's effectiveness would enable the development of adaptive components in retrieval systems [3, 36]. QPP methods can be classified according to the time of their prediction, either **pre-retrieval** (before the retrieval stage), or, **post-retrieval** (after one or more retrieval stages). The key difference between both classes is the amount and type of information that the methods use in the estimation of query quality. The key difference between each specific QPP method is the intuition underlying the method. Interestingly, while most QPP methods have been motivated and developed based on how a user might rate a query, these intuitions have never been empirically validated.

In this paper, we examine eight pre-retrieval and five post-retrieval predictors over a variety of parameter settings to determine whether the intuitions from which they are derived are in line with how users rate queries.

2.1.1 Pre-Retrieval

Pre-retrieval QPP methods estimate the performance of a query without considering the ranked list of results returned by a retrieval system in response to the query. Such methods generally exploit one of four heuristics [12] when making their prediction: *specificity*, *ambiguity*, *term relatedness* or *ranking sensitivity*. Specificity based predictors exploit collections statistics such as the inverse term or document frequencies. The intuition is that if a query has a higher specificity then it is likely to perform better than a query with low specificity [14, 23, 38]. Ambiguity based predictors either rely on external semantic sources such as WordNet [20] or on the clustering of the corpus documents [15] to determine the number of possible senses (clusters) associated with a term. Query terms that always appear in the same or similar senses or contexts across all documents, are considered unambiguous and thus better performing than queries with highly ambiguous terms. Predictors that utilize the relatedness of terms examine the relationship between query terms; highly related query terms indicate a well formed query which is likely to be successful [12]. Finally, ranking sensitivity based predictors exploit the potential sensitivity of the result ranking by predicting how easy it will be for the retrieval approach to rank the documents containing the query terms [38]. If the distribution of query terms is uniform across a large set of documents, the retrieval system is assumed to have difficulties ranking those documents and the query is considered to be of low quality.

2.1.2 Post-Retrieval

Post-retrieval predictors are employed after retrieving one or more ranked lists of results. The strategies employed are manifold. For example, a comparison between the ranked list and the corpus [8] yields a homogeneity score: the more homogeneous the top retrieved documents, the better the estimated quality of the query. Perturbing the query terms and subsequently comparing the generated ranked lists of results with respect to their overlap was evaluated in [31, 36, 40]. It was found that the higher the overlap, the less the results were influenced by a change in query term weighting and thus the better the estimated quality of the query. A similar strategy is the perturbation of documents in the ranked list of results retrieved in response to the initial query in order to determine the list's stability [31, 39]; the more stable the ranked list the better the quality of the search results (and thus the query) is assumed to be. Relying on the distribution of retrieval scores assigned to the documents in the result list has also been explored as an avenue for predicting query performance [24]. Finally, the reliance on a variety of retrieval approaches to form predictions based on the diversity of the returned documents [9] has also been proposed; the more the different retrieval methods agree on the top retrieved documents, the better the query quality.

2.1.3 QPP Evaluation

The standard methodology of evaluating QPP methods is based on comparing predicted performance scores with actual system performance. First, a QPP method computes a predicted performance score for each query of a set of topics. A system effectiveness metric, such as average precision, is also calculated for each query and a chosen retrieval approach. The correlation is then computed between the QPP-based predicted performance scores and the actual retrieval effectiveness scores. It is assumed that a higher correlation coefficient means a better QPP method[2].

[2]Most often, the linear correlation coefficient r and the rank correlation coefficient Kendall's Tau τ are reported. Both correlation coefficients lie in the interval $[-1, 1]$; a correlation close to zero indicates a lack of relationship (in the case of r a lack of a linear relationship) between the variables, while a coefficient close to ± 1 indicates a very strong relationship.

However, there has been related work examining and perhaps challenging the relationship between system-centered retrieval effectiveness measures and *actual* user performance in traditional IR settings [1, 16, 28, 30]. This research has shown consistently that traditional system-centered performance measures do not correlate well with user performance. Specifically, it has been found that when users use systems that are considered "good" according to a system-centered evaluation framework, they perform no differently than when using systems that are considered "bad" [16]. Recently, it was shown that users adapt their search behaviors to compensate for poorly performing systems and, with a little more time, are able to perform just as well as those who use a better system [25]. Overall, the findings from this body of research pose a fundamental question about the transferability of experimental results that are obtained using traditional system-centered evaluation frameworks to user scenarios. Given that QPP methods have mainly been developed within the context of a system-centered evaluation framework, it is important to examine QPP methods in a user-centered evaluation as well. Only few attempts have been made in this direction. In previous work they considered how QPP methods relate to human judgments of query quality along two lines:

1. user ratings vs. system performance, and
2. inferred user ratings vs. system performance.

Of the first line, in an experiment in the late 1990's [32], a number of IR researchers were asked to classify TREC topics as either *easy*, *medium* or *hard* for a newswire corpus they were familiar with. The researchers were given the TREC topic statements, though not the search results. It was found that they were largely unable to predict the topics' quality correctly and, surprisingly, they could not agree among themselves on how to classify the topics. Of the second line, in [29, 37] initial experiments were performed that compared a user based measure (the median time to find a relevant document) with the post-retrieval QPP method *Clarity Score* [8] and a range of pre-retrieval QPP methods [37]. In [29], no significant correlation was found for *Clarity Score*, while in [37], the best pre-retrieval predictor achieved a Kendall's Tau rank correlation of $\tau = 0.2$. However, these experiments were conducted in limited contexts, i.e., IR researchers on a small data set [32] or using time as an *implicit* user rating of query quality [29, 37]. We examine a wider range of QPP methods than previously, we consider explicit (instead of implicit) ratings of query quality by users and furthermore we investigate two different tasks, on the topic level as well as on the query suggestions level.

2.2 Term and Query Suggestions

The work on term and query suggestions is related to interactive query expansion, query substitution, and query completion. All of these approaches strive to assist users with query formulation and refinement. One can distinguish between suggestions of terms, phrases or whole queries. In the latter case the advantage is the preserved lexical coherence, while the suggestion of single terms might make sense from a statistical corpus-based point of view only.

User studies have shown that users prefer to be in control of query expansion terms, but that in many cases the use of suggested terms does not result in improved performance [19, 21, 22]. Such findings are often contrary to results obtained using system-centered evaluations of query expansion techniques which have generally shown that automatic query expansion is effective [22]. For example, even though users in [21] rated interactive query expansion better than automatic query expansion, there were no differences between the actual effectiveness of the searches. Users in Belkin et al.'s study [6] were also positive about the potential of term suggestions but did not necessarily use, or perform better with, this feature.

Several studies have been conducted to better understand potential reasons for these differences [6, 10, 11, 22]. Great variability was found by Ruthven [22] in users' abilities to identify good expansion terms, with users identifying 32-73% of the good terms. Ruthven [22] also found that users often identified terms that had a high collection frequency as being good expansion terms, even though from a system's standpoint these terms are unlikely to be useful. Certainly the statistical information is important for the system, but this information is not available to users and even if it was, it might not be useful to them. Given that many QPP methods rely on collection statistics similar to those used to identify term suggestions, it may be that the user's and system's ideas about what constitutes an effective query differ as well. In this paper, we examine whether the QPP methods that rely upon collection statistics (i.e., document frequency and collection frequency) provide a better explanation of user ratings than other QPP methods, if at all.

A number of studies have also been conducted concerning query suggestions [4, 5, 18, 26, 27, 34, 35]. Query suggestions address one of the problems identified with term suggestions in that queries are larger semantic units that are lexically more coherent than single terms. This may better assist users in determining the potential usefulness of suggestions. This is a key point since the usefulness of suggestions depends on the user identifying and using the best suggestions among a set. Systems produce many different kinds of suggestions with varying quality. Ultimately, the benefit of suggestion techniques lies in the user's ability to make predictions about the usefulness of the suggestions. This prediction, we note, occurs after the system has made its own predictions, and these query suggestions, like terms suggestions, are generated using a number of different methods. Some techniques rely on matching the current query with previously submitted queries, while other techniques attempt to automatically construct queries. Kelly et al. [18] used traditional term suggestion techniques to identify terms which were then combined with user queries in various ways to generate synthetic query suggestions. The authors compared these synthetic query suggestions to query suggestions created by other users, as well as to simple term suggestions. Overall, users selected more human-generated query suggestions, and many commented about the poor quality of the synthetic suggestions and suggested terms, which again indicates that terms predicted as useful by the system may differ from those found useful by users. This suggests that users might agree more between themselves about what makes a good query, than with, and in contrast to, QPP methods.

In summary, there is a shortfall in research examining the link between users and systems with regards to how they rate the quality of queries and predict query performance.

3. EXPERIMENTAL DETAILS

To determine whether predictions about a query's quality by users are similar to the predictions made by QPP methods, we have performed an analysis which consists of ratings and predictions from: (i) a user study which obtained query ratings pre-retrieval, at both the *topic level* and the *query suggestions level*, (ii) a secondary analysis of data from a previous user study which obtained query ratings post-retrieval at the *query suggestions level*, and (iii) corresponding predictions by pre-retrieval and post-retrieval QPP methods. The experiments were performed on two different TREC Test Collections. The results from these experiments enabled us to examine the following operational research questions:

- What is the relationship between predictions of query quality made by QPP methods and the ratings of performance made by users?
- Which QPP methods (if any) correlate significantly with users' perceptions of query quality?
- How well do QPP methods and users distinguish between well and poorly performing queries and query suggestions respectively?

In the *topic level* setup, we examined how well users are able to predict the performance of a query given the textual description of its underlying information need (the TREC topic description). That is, for each information need a single query was presented for rating. Since in our study the human assessors did not have access to the search results of these queries they acted as human pre-retrieval predictors.

In the second setup, at the *query suggestions level*, we evaluated the ability of users to judge the quality of query suggestions. For each information need/topic, the users were presented with eight different query suggestions to rate. These ratings were obtained from two different user studies, in one, pre-retrieval ratings by users were obtained based on the query and the information need (the TREC topic description), and in the other post-retrieval ratings by users were obtained based on the query, information need, and interaction with the search results.

For both setups we also obtained corresponding ratings from eight pre-retrieval and five post-retrieval QPP methods. The topic level experiment was conducted on the TREC ClueWeb09 (category B) collection, while the query suggestions level experiment was performed using the TREC Aquaint collection. An overview of the experimental setups for each collection and task level is presented in Tab. 1. The remainder of this section details the QPP methods used, and then the two user studies conducted.

	Experimental Setup			
Level:	**Topic**		**Query Suggestions**	
Collection:	ClueWeb09		Aquaint	
Prediction Type:	Pre	Post	Pre	Post
User/System Predictions				
User Study 1 (US1)	✓		✓	
User Study 2 (US2)				✓
8 Pre-Retr. QPP Methods	✓		✓	
5 Post-Retr. QPP Methods		✓		✓
Results in:	Section 4.1		Section 4.2	

Table 1: Overview of experimental conditions.

3.1 Query Performance Predictors

In our analysis, we relied upon a variety of well-known prediction algorithms there were applied to both the topic level and query suggestions level for both collections. Due to space constraints, we only briefly describe each method used and where appropriate the parameter settings used. The **pre-retrieval predictors** evaluated were:

- Average (*AvIDF*) and Maximum (*MaxIDF*) Inverse Document Frequency [8, 23],
- Average Query Word Length (*AvQL*) [20],
- Average (*AvSCQ*) and Summed (*SumSCQ*) Similarity of Collection and Query [38],
- Average (*AvVAR*) and Summed (*SumVAR*) Term Weight Variability [38], and,
- Average Pointwise Mutual Information (*AvPMI*) [12].

AvIDF, *MaxIDF*, *AvSCQ* and *SumSCQ* belong to the class of specificity-based predictors, which rely on term and document frequencies of the query terms in the corpus to derive a predicted quality score. *AvQL* is part of this category as well, but considers only the average number of characters in a query - the assumption being that longer terms are less common in a corpus and thus more specific. In contrast, *AvVAR* and *SumVAR* exploit the distribution of TF.IDF based term weights in the corpus and thus belong to the ranking sensitivity category. Finally, the relatedness between query terms is expressed by *AvPMI*, where a higher relatedness between query terms indicates a better query quality. Please note, that all of these QPP methods are parameter-free.

The **post-retrieval predictors**, i.e. those that rely on one or more retrieved result lists, evaluated were as follows:

- *Ranking Robustness* [39],
- *Spatial Autocorrelation*[3] [9],
- *Query Feedback* [40],
- *Clarity Score* [8, 13], and,
- *Query Commitment* [24].

These methods were chosen due to their state-of-the-art performance and the diversity of approaches they represent. In the *Ranking Robustness* [39] approach, the top m retrieved documents of the initial search are perturbed by adding or removing terms. The perturbed documents are then ranked based on the original query and retrieval approach. The higher the correlation between the original and perturbed result list, the higher the predicted query quality. We evaluated this approach for $m = \{10, 50, 100, 250, 500, 1000, 2500, 5000\}$.

In the *Spatial Autocorrelation* [9] method, a document's retrieval score is replaced by the weighted sum of retrieval scores of its k most similar documents in the retrieved result list as determined by TF.IDF. The linear correlation coefficient between the original document scores and the perturbed document scores form the predicted query quality score. The parameter k was varied between 2, 5, 10 and 15.

In contrast to document (score) perturbation, *Query Feedback* [40] is based on query perturbations: from the originally retrieved top m ranked documents, a new, perturbed query is generated consisting of the n most discriminative

[3]Referred to as $\rho(\mathbf{y}, \tilde{\mathbf{y}})$ in [9].

terms. A second ranked list is retrieved based on the perturbed query and the overlap between the two lists is utilized as query quality score. The lower the overlap between the two lists, the lower the predicted query quality. We evaluated all possible parameter combinations with $m = \{10, 20, 30, 40, 50, 100\}$ and $n = \{10, 20, 30\}$.

Clarity Score [8] is based on the intuition that the top m ranked documents of an unambiguous query will be topically cohesive and terms particular to the topic will appear with high frequency. Thus, the higher the difference between the term distribution of the top retrieved documents and the term distribution of the corpus, the higher the predicted query performance. We experimented with values of $m = \{10, 50, 100, 250, 500, 750, 1000\}$. An adaptation of Clarity Score, which ignores terms in the calculation that appear in more than $\frac{1}{k}$th of the corpus was introduced in [13]. We evaluated $k = \{1, 10, 100\}$.

Lastly, *Query Commitment* [24] determines the standard deviation σ of the retrieval scores of the top m retrieved documents, possibly normalized by a query-dependent corpus statistic. The higher σ, the higher the difference in retrieval scores in the result list, indicating a few top ranked documents with a high query commitment and thus high quality is predicted. We experimented with $m = \{10, 50, 100, 250, 500, 750, 1000, 2500, 5000, 10000\}$.

3.2 User Studies

Two user studies provided the human ratings of query quality at the topic level and the query suggestions level, as indicated by Tab. 1. Details of each study are described below.

3.2.1 US1: Pre-Retrieval Assessor Ratings

Following on from the previous experiments [29, 32, 37], we performed a similar experiment at the *topic level* using the most recent TREC test corpus: ClueWeb09 (category B) [7], a 50 million document crawl of the Web from 2009. We utilized the fifty topics of the TREC 2009 Web *ad hoc* retrieval task[4] which consist of a *query* part (to be submitted to the IR system) and a *description* (the information need). E.g., the topic wt09-3 consists of the query *"getting organized"* and the description *"Find tips, resources, supplies for getting organized and reducing clutter"*.

We provided assessors with the queries and descriptions and instructed them to judge, on a scale from 1 (poor quality query) to 5 (high quality query), the queries according to "what you expect the search results to be, if you would submit these queries to a Web search engine (keeping in mind the actual information need)". The presentation order of the queries was randomized across assessors. Note, that the queries were not actually submitted to a search engine.

In a second experiment, this time on the *query suggestions level*, the same assessors were asked to judge eight query suggestions for each of four topics taken from the topics of the TREC Robust track [33] which were based on the Aquaint corpus of news stories. The particular query suggestions used were obtained from user study US2, described in Section 3.2.2. The assessors were presented the topic descriptions and the query suggestions, but not the result list, thus providing sets of pre-retrieval query quality ratings for the query suggestions level.

[4]One topic (wt09-20) has no relevant documents and was ignored.

Eighteen users, recruited via email solicitation from two university research groups (Databases, Human Media Interaction) participated in the experiments. Eleven participants were male. Lemur[5] was utilized as the underlying retrieval system. The document indices were Porter stemmed and stopwords were removed. The KL-divergence based retrieval model was relied upon to calculate the system-oriented ground truth with respect to TREC relevance judgments (Dirichlet smoothing with $\mu = 1000$). The retrieval effectiveness was measured in precision at 30 documents (P@30) and average precision (AP) for the Aquaint corpus. In the case of ClueWeb09, AP and P@30 were estimated according to [2], which is the evaluation measure/procedure for this corpus at TREC. To denote the difference between these measures, we refer to the ClueWeb09 measures as *estimated* AP and *estimated* P@30 respectively.

3.2.2 US2: Post-Retrieval Assessor Ratings

Data for this second study, which was performed on the *query suggestions level*, was generated from a user study performed previously where the use of query suggestions was the focus [17]. For our experiments here, we have performed a secondary analysis of the data. The basic goal of the previous study was to examine subjects' selection of query suggestions.

This study involved subjects engaging in search using an experimental IR search application, where Lemur was used for indexing and retrieval. As part of this experience, the subjects evaluated the quality of query suggestions that were presented to them after they had ended their search session for a search topic.

Four of the eight presented query suggestions were high quality queries and four were low quality queries, a classification that was performed by examining the number of relevant documents retrieved in the top 20 results. An example is TREC topic 354 from the TREC Robust track whose description is *"Identify instances where a journalist has been put at risk (e.g., killed, arrested or taken hostage) in the performance of his work. Any document identifying an instance where a journalist or correspondent has been killed, arrested or taken hostage in the performance of his work is relevant."*; two high quality query suggestions in this case were *"journalist killed"* and *"journalists arrested for work"* wheras two poorly performing suggestions were *"reporter killed"* and *"journalist at risk reporting danger"*. The average number of terms per query suggestion was 3.19. Note, the query suggestions for User Study 1 and User Study 2 were the same. Since the participants of this study rated the query suggestions *after* they completed their search task, they acted as human post-retrieval predictors. The participants of this study were twenty-three university students who responded to a campus-wide email solicitation for research subjects. Most of them were female ($n = 16$) and their average age was 21 years. Students' majors varied across the humanities, social sciences and sciences.

4. ANALYSIS

We first report the results on the *topic level* experiments in Sec. 4.1. Then, in Sec. 4.2 we give an overview of the *query suggestions level* experiments. In both instances, we analyze: (i) the human assessor ratings of query quality, (ii)

[5]The Lemur Toolkit, `http://www.lemurproject.org/`

the QPP methods' predictions of query quality, and, (iii) the relationship between assessor ratings and the QPP methods' predictions.

4.1 Topic Level Experiment

Recall, that in this setup, the assessors and QPP methods predict the quality of a single query per topic or information need. The assessor ratings were collected in User Study 1 which focused on the pre-retrieval setting.

4.1.1 Assessor Ratings of Query Quality

The ratings of query quality by the 18 study participants varied considerably, leading to a relatively low inter-rater agreement. When comparing all possible pairs of participants, we observed a maximum linearly weighted Cohen's Kappa of $\kappa = 0.54$, which is a moderate agreement. The median agreement between all pairs reached $\kappa = 0.36$, while the minimum amounted to $\kappa = 0.12$, indicating low agreement. These findings echo those found in [32], where a low agreement among assessors was also noted.

But, how did the assessors fare in identifying queries that performed well and those that performed poorly according to system effectiveness measures? To investigate this, we split the 49 queries into four partitions with ten queries and one with nine based on system performance. This was done for both estimated AP and P@30, to provide two rankings of query quality according to system performance (Tab. 2, columns 2&3). In both instances, the top ten performing queries were in partition one, the next best performing queries were in partition two, and so on. For each measure and each partition, we then averaged all observed assessor ratings for the queries within the partition. Columns 4&5 in Tab. 2 show the average user rating and the standard deviation σ for each partition and measure. The trend suggests that the assessors rated poorly performing queries lower than the highly performing queries. When we consider the results for estimated AP (P@30 is similar), particularly stark is the contrast between the best and worst partition: while the average rating for queries of the best partition is 3.87 (out of a max. of 5.0), the worst partition is assigned an average of 2.51 (out of a min. of 1.0). This suggests that although the assessors do not agree to a high degree with each other on the quality ratings, on average, they are able to distinguish good from bad queries at the topic level.

To evaluate the relationship between assessor ratings and system performance in more detail, we computed the correlation between assessor ratings and retrieval effectiveness, following the methodology that is used to evaluate QPP methods. Here, we report the rank correlation coefficient Kendall's Tau, a standard QPP evaluation measure. The worst correlated assessor obtains a correlation of $\tau = 0.17$ with estimated AP ($\tau = 0.20$ with P@30), the median correlation coefficients are $\tau = 0.31$ for AP and $\tau = 0.35$ for P@30 respectively (both statistically significant at $p < 0.01$). The most highly correlated assessor reaches a correlation of $\tau = 0.47$ with estimated AP and $\tau = 0.45$ with P@30 (both statistically significant at $p < 0.01$). This result shows that the assessors' ability to rate the quality of queries correctly varies significantly, despite the fact that the assessors all have a comparable educational and search-engine-usage background. It also indicates that although, on average, assessors could rate the quality of queries given the different bands of query quality (as shown in Tab. 2), the assessors had much more difficulty precisely rating the quality of queries.

4.1.2 QPP Methods' Predictions

As mentioned, QPP methods are usually evaluated by reporting their correlation with retrieval effectiveness measures. For completeness, we report these results in Tab. 3, columns 2&3. In the case of the post-retrieval QPP methods, we report the results of the best performing parameter settings only. Note that at the topic level for this particular corpus, the pre-retrieval QPP methods *SumSCQ* and *SumVAR* achieve the highest correlation and thus outperform the more complex post-retrieval QPP methods. This is in contrast to previous findings where in older test corpora it is the more complex QPP methods that obtain higher correlations. We suspect that in the case of ClueWeb09, which was derived from a recent crawl of the Web, relying on document content as the post-retrieval QPP methods do, can also be a disadvantage as nowadays Web pages do not only contain informational content, but also a large amount of non-informative content (e.g., navigational elements, advertisements, spam, etc.), which may adversely affect the abilities of these predictors.

QPP Methods	System		Assessor Ratings		
	AP	P@30	Min.	Med.	Max.
Pre-Retrieval					
AvIDF	0.28⋆	0.15	−0.14	0.05	0.23
MaxIDF	0.35⋆	0.19	−0.09	0.10	0.29⋆
AvQL	0.21	0.04	−0.01	0.12	0.24
AvSCQ	0.27⋆	0.14	−0.13	0.04	0.24
SumSCQ	0.39⋆	0.35⋆	0.20	0.31⋆	0.49⋆
AvVAR	0.30⋆	0.19	−0.06	0.08	0.24
SumVAR	0.42⋆	0.38⋆	0.17	0.30⋆	0.43⋆
AvPMI	0.25	0.23	0.09	0.23	0.41⋆
Post-Retrieval					
Robustness $m = 1000$	0.08	0.07	−0.08	0.02	0.16
Spatial Autocorr. $k = 2$	0.15	0.20	−0.10	0.04	0.13
Query Feedback $m = 20, n = 10$	0.37⋆	0.29⋆	0.12	0.29	0.44⋆
Clarity Score $m = 100, k = 100$	0.27⋆	0.18	−0.10	0.04	0.19
Query Commit. $m = 5000$	0.28⋆	0.13	−0.13	0.01	0.17

Table 3: Topic level experiment: Kendall's Tau correlation coefficients between QPP methods and system effectiveness (columns 2&3). The correlation between QPP methods and assessor ratings (minimum, median and maximum Kendall's Tau) are listed in columns 4-6. Significant correlations ($p < 0.01$) are marked with a star.

4.1.3 Assessor Ratings vs. QPP Methods

So far we have seen that neither assessor ratings nor QPP methods correlate highly with system effectiveness, at the topic level. Given that there is quite a mismatch with system performance, it may be the case that there is higher agreement between the assessors and QPP methods. So now, we turn our attention to the focus of this paper, and determine whether at the topic level the predictions of query performance by QPP methods fall in line with the quality ratings made by the assessors, and consider the correlation between them. The results are reported in Tab. 3, columns 4-6. Due

Query Partitions	Avg. System Performance wrt.		Avg. Assessor Ratings wrt.	
	estimated AP	estimated P@30	estimated AP	estimated P@30
Queries ranked 1-10	0.414	0.629	3.87 ($\sigma = 1.07$)	4.00 ($\sigma = 1.01$)
Queries ranked 11-20	0.298	0.470	3.72 ($\sigma = 1.09$)	3.53 ($\sigma = 1.20$)
Queries ranked 21-30	0.099	0.272	3.24 ($\sigma = 1.37$)	3.31 ($\sigma = 1.29$)
Queries ranked 31-40	0.032	0.133	2.79 ($\sigma = 1.20$)	2.89 ($\sigma = 1.33$)
Queries ranked 41-49	0.005	0.038	2.51 ($\sigma = 1.48$)	2.40 ($\sigma = 1.34$)

Table 2: Topic level experiment: Average system performance of the query partitions based on estimated AP and P@30 respectively (columns 2&3); average (std. deviation σ) assessor ratings of the query partitions based on estimated AP and P@30 respectively (columns 4&5).

to the, at best, moderate level of agreement between the assessors we resort to reporting the minimum, the median and the maximum correlation between the assessor ratings and the QPP methods. We observe the highest min., med. and max. correlation between assessor ratings and *SumSCQ* [38]. This predictor combines the collection term frequency and inverse document frequency. It is summed over all query terms of a query $Q = \{q_1, ..., q_m\}$:

$$SumSCQ = \sum_{i=1}^{m} (1 + \ln(cf(q_i))) \ln\left(1 + \frac{doccount}{df(q_i)}\right)$$

Here, *doccount* is the number of documents in the corpus, *df* is the document frequency and *cf* is the collection term frequency. Zhao et al. [38] argue that a query, which is similar to the corpus as a whole is easier to retrieve documents for, since the similarity is an indicator of whether documents answering the information need appear in the corpus. Since the score assigned to a query is proportional to collection term frequency and inverse document frequency of terms, the terms that appear in few documents many times are favored. Those terms are highly specific, as they occur in relatively few documents, while at the same time they occur often enough to be important to the query.

It is also worth mentioning, that *SumSCQ* and *SumVAR* achieve significant correlations with most assessors (the median correlation is significant), while the remaining 11 QPP methods only sometimes obtain a significant correlation with few assessors at best. The post-retrieval QPP methods, which are reported with their best parameter settings, apart from *Query Feedback* all perform poorly, resulting in no significant correlation with any human assessor.

4.2 Query Suggestions Level Experiments

For the *query suggestions level* experiments, we relied on the data from the two user studies; in US1 (Sec. 3.2.1) the assessors were asked to rate query suggestions without access to the search results (i.e., pre-retrieval) while in US2 (Sec. 3.2.2) the assessors actually performed different search tasks and rated the query suggestions after completing each search session (i.e., post-retrieval). Recall, that at the query suggestions level, assessors were given eight query suggestions to rate per topic - four high quality suggestions and four low quality suggestions.

4.2.1 Assessor Ratings of Query Quality

In the **pre-retrieval** setup, when we averaged the user ratings over all high and low quality query suggestions respectively, the high quality query suggestions received an average rating of 3.74 ($\sigma = 1.02$) while the low quality suggestions were rated on average with 2.83 ($\sigma = 1.23$). The inter-rater agreement between any of the assessors was, at best, $\kappa = 0.55$, while the median was $\kappa = 0.25$ and the minimum $\kappa = -0.05$. When comparing these numbers to the topic level experiments (the same set of assessors), we note that while on average the group can distinguish good from bad query suggestions in terms of system effectiveness, when it comes to the agreement between assessors at the query suggestion level there was less agreement between assessors - in the topic level experiments the median agreement reached $\kappa = 0.36$. This may suggest that rating query suggestions is a more difficult task.

Query Suggestions		Query Quality	
		High	Low
US1: Pre-Retr.	Avg. Rating	3.74 (1.02)	2.83 (1.23)
US2: Post-Retr.	Avg. Rating	3.00 (1.15)	2.80 (1.18)
	#Suggestions Used	170 of 368	143 of 368
	Avg. Rating Used	3.02 (1.16)	2.68 (1.24)
	Avg. Rating Unused	2.97 (1.13)	2.87 (1.12)

Table 4: Query suggestions level experiment: rating overview of the high and low quality query suggestions the study participants were offered as well as those that the participants used/not used.

In the **post-retrieval** setup, the assessors gave the high quality queries an average rating of 3.00 ($\sigma = 1.15$), and the low quality queries on average were rated as 2.80 ($\sigma = 1.18$). While, on average, the subjects from US2 tended to rate the effective query suggestions higher, it was not to the same extent as the assessors in US1. Along with the average ratings for the high and low quality query suggestions, Tab. 4 also reports the total number of times the two groups of query suggestions were issued by the users in US2, out of the possible number of suggestions for each type. The participants issued significantly more high quality query suggestions than low quality suggestions ($\chi 2(1, 736) = 5.95$, $p = 0.015$) and rated high quality queries significantly higher than low quality queries ($t(734) = 2.38$, $p = 0.017$). While the difference was significant, it was lower than in the pre-retrieval setup. This suggests that the influence of the interaction with the search results and issuing only a subset of the possible query suggestions may affect the ratings of the suggestions. In order to investigate whether a confirmation bias exists in the ratings, that is, whether a subject selects a query suggestion (good or bad) and then feels the need to justify her decision and rate the suggestion a bit higher than suggestions that she did not select, we investigated the average rating of query suggestions that were used and not used respectively by the subjects. The results are also shown in Tab. 4: while the difference in rating is very small for the high quality queries between the used and unused suggestions, in the case of poorly performing suggestions, larger differences appear: query suggestions that were issued re-

ceive a lower average rating than suggestions that were not used.

The inter-rating agreement of assessors in the post-retrieval setup was lower than in the pre-retrieval setup. The median agreement among these participants was $\kappa = 0.01$ and the minimum and maximum were $\kappa = -0.26$ and $\kappa = 0.33$ respectively. This again suggests that the queries each participant issued, and the different search results that the participants interacted with, also leads to less agreement. This is presumably because each participant's state of knowledge has changed in different ways. Whereas in the pre-retrieval setting the assessors only have their base knowledge and the topic description, so the information that they have available is common to all assessors, which may be the reason for the higher levels of agreement between them.

4.2.2 QPP Methods' Predictions

The correlations achieved by the QPP methods at the query suggestions level are reported in Tab. 5, columns 3&4. We found the post-retrieval QPP methods to exhibit a higher correlation with system performance than the pre-retrieval predictors. Specifically, *Query Feedback*, *Clarity Score* and *Query Commitment* were all significantly correlated with AP and the former two with P@30 as well. On the other hand, of the pre-retrieval predictors only *AvQL* was significantly correlated with P@30. This is an interesting finding because it shows that when comparing queries for the same topic, pre-retrieval predictors are not as indicative of performance.

4.2.3 Assessor Ratings vs. QPP Methods

We now turn our attention again to the main research question and consider the relationship between the assessor ratings and QPP methods' predictions. At the query suggestions level we can examine this relationship with respect to (i) pre-retrieval ratings (with assessors from US1), and (ii) post-retrieval ratings (with the assessors from US2). Columns 5-10 of Tab. 5 report the correlations between the predictions made by QPP methods and the ratings of query quality made by the assessors. As earlier, we report the minimum, median and maximum correlation across all assessors due to the low inter-rater agreement. Several interesting observations can be made.

In the pre-retrieval setup (US1), we observed that the correlation between assessors and QPP methods tends to be positive overall, however, no QPP methd is significantly correlated with the majority of assessors, that is, the median correlation is not significant. The majority of QPP methods, both pre- and post-retrieval, were significantly correlated with the most correlated assessor. Similarly to the topic level experiments, *SumSCQ* and *SumVAR* performed very well with respect to the other QPP methods. This suggests that these predictors are the most indicative of assessor ratings, although the relationship appears stronger at the topic level than at the query suggestions level. In contrast to the topic level experiments, the majority of evaluated post-retrieval QPP methods, specifically *Robustness*, *Spatial Autocorrelation* and *Query Feedback*, lead to a significant correlation with the most correlated assessor.

In the post-retrieval setup (US2), the results were markedly different. The correlation between assessors and QPP methods varied considerably, ranging from negative to positive correlations. The median correlation for the post-retrieval assessor with any of the QPP methods was close to, or around, zero, indicating that there was no relationship between them. In fact, some assessors were quite negatively correlated with QPP methods (see minimum correlations) which ranged from -0.36 to -0.14. The only significant correlation ($\tau = 0.48$) can be reported for *Query Feedback*. These results suggest that the post-retrieval assessments by participants are very varied, and with respect to QPP methods are not consistently related in one way or another. It appears that the interaction with results and usage of the system has a considerable effect on the subjects ratings of query quality.

5. DISCUSSION AND CONCLUSIONS

We have performed and analyzed two empirical studies comparing the predictions of query quality made by automatic QPP methods with the predictions made by human assessors. We have conducted a more comprehensive analysis than previous work by examining the relationship of assessors with thirteen pre- and post-retrieval QPP methods, at both the topic level and the query suggestions level. From this work, our main findings relating to our operational research questions outlined in Sec. 3, indicate that:

- The correlation between the predictions derived from QPP methods and the predictions obtained from human assessors was quite weak at both the topic and the query suggestions level. We also have to conclude that overall QPP methods are not representative of how assessors rate the quality of queries. There were, however, some notable exceptions (see next point).
- When quality ratings were obtained pre-retrieval for both the topic and query suggestions level, the pre-retrieval QPP methods *SumSCQ* and *SumVar* as well as the post-retrieval QPP method *Query Feedback* (in its best parameter setting) did exhibit moderate correlations with assessors and in most cases also significant correlations.
  - Across the different experiments, the QPP method that best reflected assessors' ratings of query quality was the pre-retrieval approach *SumSCQ*.
- Overall, and on average, assessors were able to distinguish between "good" (i.e., effective) and "bad" (i.e., ineffective) queries at the topic level and the query suggestions level based on the system effectiveness measures (Tab. 2 and 4).

While the findings from this work are partially in tune with previous research (e.g., Turpin & Hersh [29] found that *Clarity Score*, a commonly employed post-retrieval QPP approach, has no correlation with implicit user ratings of query quality), we have teased out more precisely which QPP methods exhibit a relationship with explicit user ratings of query quality. In contrast to [32], we have found that, on average, quality ratings of queries tended to be in line with system performance at both the topic and the query suggestions level. However, query quality ratings obtained post-retrieval did not emphasize the difference in quality as well as those ratings obtained pre-retrieval from human assessors. Unfortunately, we are unable to definitively provide reasons for this difference. This is due to the number of experimental

QPP Methods		System		Assessor Ratings					
				US1: Pre-Retrieval			US2: Post-Retrieval		
	Parameter(s)	AP	P@30	Min.	Med.	Max.	Min.	Med.	Max.
Pre-Retrieval									
AvIDF		0.14	0.06	−0.01	0.15	0.32	−0.23	0.01	0.28
MaxIDF		0.19	0.06	0.01	0.16	0.36⋆	−0.22	0.00	0.21
AvQL		0.28	0.37⋆	−0.02	0.19	0.38⋆	−0.16	0.15	0.31
AvSCQ		0.21	0.15	−0.03	0.13	0.28	−0.18	−0.07	0.24
SumSCQ		0.20	0.16	−0.08	0.24	0.44⋆	−0.22	−0.02	0.31
AvVAR		0.14	0.05	−0.01	0.12	0.30	−0.29	−0.02	0.21
SumVAR		0.18	0.06	−0.02	0.21	0.44⋆	−0.24	0.00	0.20
AvPMI		0.25	0.18	0.06	0.23	0.35⋆	−0.14	0.05	0.27
Post-Retrieval									
Robustness	$m = 250$	0.32	0.28	0.12	0.28	0.44⋆	−0.18	0.03	0.36
Spat. Autocorr.	$k = 2$	0.27	0.21	0.07	0.28	0.43⋆	−0.18	0.04	0.32
Query Feedback	$m = 50, n = 20$	0.51⋆	0.43⋆	0.10	0.29	0.45⋆	−0.36	0.01	0.48⋆
Clarity Score	$m = 10, k = 1$	0.41⋆	0.41⋆	0.00	0.08	0.20	−0.24	−0.04	0.22
Query Commit.	$m = 500$	0.36⋆	0.29	0.02	0.18	0.29	−0.20	0.03	0.31

Table 5: Query suggestions level experiment: Kendall's Tau correlation coefficients between QPP methods and system effectiveness are shown in columns 3&4. The correlation between QPP methods and the US1/US2 assessor ratings (minimum, median and maximum Kendall's Tau) are listed in columns 5-10. Significant correlations ($p < 0.01$) are marked with a star. For the post-retrieval QPP methods the best performing parameter settings are shown.

variables and the limitations of the studies conducted here. Nonetheless, we have made substantial progress in determining the relationship between automatic QPP methods and user ratings of query performance. And, we have identified a number of factors which appear to influence the ratings of queries. We discuss them in turn below:

User Background In US1, the assessors may have had more system knowledge as they were largely computer science post-graduates, where as in US2 the assessors were undergraduates mainly from humanities. These different backgrounds may be the source of the differences observed given the ratings of query suggestions.

Information Available While it is likely that the background of the assessors may influence their ratings, the information provided to assessors to rate the queries is probably a source of greater variation. Depending on whether the assessor sees the query, or the query and the underlying information need, along with the search results, it is likely to affect the assessor's perception of the query's quality. In particular, the interaction with the system and the engagement with the information is also likely to influence his ratings (see below).

Interaction The interaction with the search system by the subjects as observed in US2 clearly impacted the ratings of query quality and is very likely to be responsible for the differences we observed between the groups. We found a considerable difference on the query suggestions level between the assessors who predicted the query quality before the retrieval stage and after completing the search task. Assessors rating the suggestions pre-retrieval, agreed with each other to a higher degree than assessors who rated the suggestions post-retrieval. We posit that during the course of searching the cognitive states of the assessors would have changed in various ways depending on their interaction. This divergence from the initial state of query and information need is likely to account for the greater variation between user agreement and quality ratings.

Obtaining Ratings Many difficulties are presented when obtaining ratings for queries. This is especially the case in the post-retrieval setup, where there are a number of potential variables which may influence the ratings. The main question is when should the query quality ratings be obtained? Should one wait until the end of the search task (as we did in our experiment) when they might have problems distinguishing and remembering the different queries? Or should they be interrupted each time they issue a new query to rate the previous one? Also, at what point does it become a rating of actual performance, as opposed to a rating of predicted performance?

Topic or Query Suggestions The difference between evaluating one query per topic (topic level) or multiple queries per topic (query suggestions level) may also play a role. QPP methods are traditionally evaluated and optimized for the former, possibly a reason why the results for the topic level setup are more in line with actual user ratings. In the query suggestions level setup, the ratings are more comparative in nature, and so it might be that "ranking" suggestions is more appropriate than "rating" suggestions.

System Dependence vs. Method Independence Though not specifically considered in this study explicitly, the employed retrieval approach is also a factor when investigating the query quality ratings against system effectiveness. Pre-retrieval QPP methods predict a query's performance independent of the retrieval approach. While this is not problematic when working with TREC corpora, where the ranking functions employed are similar, it becomes an issue when moving to the Web where search engines have access to a lot of additional features such as links, click-through data, etc. Note, that since the focus of this work was on whether QPP methods make predictions in line with users this is perhaps not relevant here, but may be applicable to future work.

User Expectation With the prevalence of highly effective search engines users may be expecting a certain level of performance. This *a priori* expectation may influence the ratings of queries, and so the context of the system needs to be considered in the rating process.

In summary, we have performed an analysis comparing and relating the predictions of QPP methods to the ratings performed by users. While some valuable insights have been gained by this study, substantially more research needs to be conducted in this direction. In particular, there needs to be a concentrated effort on understanding and developing methods for the user-sided query suggestions level; as this task appears to be more important than the standard QPP prediction task in the context of developing adaptive Information Retrieval systems. In future research we would like to explore the influence of the above factors on query performance prediction and to develop more sophisticated predictive models that include the user, their state of knowledge and the retrieval system within the process.

6. REFERENCES

[1] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? In *SIGIR '05*, pages 433–440, 2005.

[2] J. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In *SIGIR '06*, pages 541–548, 2006.

[3] R. Baeza-Yates, V. Murdock, and C. Hauff. Efficiency trade-offs in two-tier web search systems. In *SIGIR '09*, pages 163–170, 2009.

[4] R. Baraglia, F. Cacheda, V. Carneiro, D. Fernandez, V. Formoso, R. Perego, and F. Silvestri. Search shortcuts: a new approach to the recommendation of queries. In *RecSys '09*, pages 77–84, 2009.

[5] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *KDD '00*, pages 407–416, 2000.

[6] N. J. Belkin, C. Cool, D. Kelly, S.-J. Lin, S. Y. Park, J. Perez-Carballo, and C. Sikora. Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. *Inf. Process. Manage.*, 37(3):403–434, 2001.

[7] C. L. Clarke, N. Craswell, and I. Soboroff. Preliminary report on the TREC 2009 Web Track. In *TREC 2009 Notebook Papers*, 2009.

[8] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *SIGIR '02*, pages 299–306, 2002.

[9] F. Diaz. Performance prediction using spatial autocorrelation. In *SIGIR '07*, pages 583–590, 2007.

[10] A. Diriye, A. Blandford, and A. Tombros. A polyrepresentational approach to interactive query expansion. In *JCDL '09*, pages 217–220, 2009.

[11] M. Hancock-Beaulieu and S. Jones. Interactive searching and interface issues in the okapi best match probabilistic retrieval system. *Interacting with Computers*, 10(3):237–248, 1998.

[12] C. Hauff, D. Hiemstra, and F. de Jong. A survey of pre-retrieval query performance predictors. In *CIKM '08*, pages 1419–1420, 2008.

[13] C. Hauff, V. Murdock, and R. Baeza-Yates. Improved query difficulty prediction for the web. In *CIKM '08*, pages 439–448, 2008.

[14] B. He and I. Ounis. Inferring query performance using pre-retrieval predictors. In *SPIRE'04*, pages 43–54, 2004.

[15] J. He, M. Larson, and M. de Rijke. Using coherence-based measures to predict query difficulty. In *ECIR'08*, pages 689–694, 2008.

[16] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson. Do batch and user evaluations give the same results? In *SIGIR '00*, pages 17–24, 2000.

[17] D. Kelly, A. Cushing, M. Dostert, X. Niu, and K. Gyllstrom. Effects of popularity and quality on the usage of query suggestions during information search. In *CHI '10*, pages 45–54, 2010.

[18] D. Kelly, K. Gyllstrom, and E. W. Bailey. A comparison of query and term suggestion features for interactive searching. In *SIGIR '09*, pages 371–378, 2009.

[19] J. Koenemann and N. J. Belkin. A case for interaction: a study of interactive information retrieval behavior and effectiveness. In *CHI '96*, pages 205–212, 1996.

[20] J. Mothe and L. Tanguy. Linguistic features to predict query difficulty - a case study on previous trec campaigns. In *SIGIR'05 Query Prediction Workshop*, 2005.

[21] Y. Nemeth, B. Shapira, and M. Taeib-Maimon. Evaluation of the real and perceived value of automatic and interactive query expansion. In *SIGIR '04*, pages 526–527, 2004.

[22] I. Ruthven. Re-examining the potential effectiveness of interactive query expansion. In *SIGIR '03*, pages 213–220, 2003.

[23] F. Scholer, H. Williams, and A. Turpin. Query association surrogates for web search. *J. Am. Soc. Inf. Sci. Technol.*, 55(7):637–650, 2004.

[24] A. Shtok, O. Kurland, and D. Carmel. Predicting Query Performance by Query-Drift Estimation. In *ICTIR '09*, pages 305–312, 2009.

[25] C. L. Smith and P. B. Kantor. User adaptation: good results from poor systems. In *SIGIR '08*, pages 147–154, 2008.

[26] B. Smyth, E. Balfe, J. Freyne, P. Briggs, M. Coyle, and O. Boydell. Exploiting query repetition and regularity in an adaptive community-based web search engine. *User Modeling and User-Adapted Interaction*, 14(5):383–423, 2005.

[27] M. Strohmaier, M. Kröll, and C. Körner. Intentional query suggestion: making user goals more explicit during search. In *WSCD '09*, pages 68–74, 2009.

[28] A. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *SIGIR '01*, pages 225–231, 2001.

[29] A. Turpin and W. Hersh. Do clarity scores for queries correlate with user performance? In *ADC '04*, pages 85–91, 2004.

[30] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *SIGIR '06*, pages 11–18, 2006.

[31] V. Vinay, I. J. Cox, N. Milic-Frayling, and K. Wood. On ranking the effectiveness of searches. In *SIGIR '06*, pages 398–404, 2006.

[32] E. Voorhees and D. Harman. Overview of the sixth text retrieval conference (trec-6). In *Proceedings of the Sixth Text REtrieval Conference*, 1997.

[33] E. M. Voorhees. Overview of the TREC 2005 Robust Retrieval Track. In *TREC 2005*, 2005.

[34] R. W. White, M. Bilenko, and S. Cucerzan. Studying the use of popular destinations to enhance web search interaction. In *SIGIR '07*, pages 159–166, 2007.

[35] R. W. White and G. Marchionini. Examining the effectiveness of real-time query expansion. *Inf. Process. Manage.*, 43(3):685–704, 2007.

[36] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow. Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In *SIGIR '05*, pages 512–519, 2005.

[37] Y. Zhao and F. Scholer. Predicting query performance for user-based search tasks. In *ADC '07*, pages 112–115, 2007.

[38] Y. Zhao, F. Scholer, and Y. Tsegay. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *ECIR '08*, pages 52–64, 2008.

[39] Y. Zhou and W. B. Croft. Ranking robustness: a novel framework to predict query performance. In *CIKM '06*, pages 567–574, 2006.

[40] Y. Zhou and W. B. Croft. Query performance prediction in web search environments. In *SIGIR '07*, pages 543–550, 2007.