

Predicting the Effectiveness of Queries and Retrieval Systems

CLAUDIA HAUFF

Chapter 3

Post-Retrieval Prediction: Clarity Score Adaptations

3.1 Introduction

The previous chapter focused on pre-retrieval prediction methods which are mostly based on the query terms' collection statistics. We now turn our attention to post-retrieval methods which derive their performance estimations from ranked lists of results retrieved in response to a query. Post-retrieval predictors are computationally more expensive than pre-retrieval approaches since at least one, or possibly more, retrieval round has to be performed before a prediction can be made. However, the drawback of increased complexity is considered worthwhile, as an increased accuracy in the predictions is expected due to the greater amount of information available. As evident in the the previous chapter, *TFIDF*, for instance, is a very poor retrieval approach for our available corpora, whereas Okapi achieves adequate retrieval effectiveness. A pre-retrieval predictor, however, assigns the same predicted score in both cases to a given query, while a post-retrieval predictor derives two distinct predictive scores as it is based on the ranked list of results.

In this chapter, we first present an overview of post-retrieval prediction methods. Then, we will focus on one post-retrieval approach in particular, namely Clarity Score, which was proposed by Cronen-Townsend et al. [45]. Clarity Score is based on the intuition that the top ranked results of an unambiguous and thus well performing query are topically cohesive, whereas the result list of an ambiguous and thus poorly performing query will cover a variety of topics. The degree of topical cohesiveness is derived from the term distribution of the top ranked documents: a homogeneous result list will lead to a distribution where terms particular to the topic appear with high frequency, while topically not homogeneous results with documents covering a variety of topics are assumed to be more similar to the collection distribution. Consider, for instance, the example query “jaguar” and its results, described in Chapter 1. The result list of the first 500 documents would be predicted to be of high quality by Clarity Score, as by far most results are concerned with the topic of cars. If, however, we were only to consider the result list up to rank ten, the query would be predicted to be somewhat ambiguous, as three documents are

concerned with the animal, while seven documents are concerned with cars. This sensitivity to the number of documents to rely on in the Clarity Score calculation is problematic. Currently, it is necessary to search exhaustively through the parameter space in order to find a reasonable setting.

In this work, we propose *Adaptive Clarity* which addresses two perceived shortcomings of Clarity Score. First, Clarity Score uses a fixed number of top-retrieved documents from which to derive the term distribution - we show that this is not optimal and propose an easy solution for a query adaptive automatic setting of this parameter. Second, the predicted performance score is the Kullback-Leibler (KL) divergence [91] between the term distribution of the top ranked results and the term distribution of the entire collection. Although all terms of the vocabulary participate in this equation, terms that have a high document frequency in the collection (and which for some reason occur uncharacteristically rarely in the top retrieved documents) add nothing to distinguish a set of homogeneous documents from the collection and thus we use a threshold of maximum document frequency to exclude those terms.

We will show in this chapter the following:

- Clarity Score in its original form is very sensitive to its parameter setting, and,
- Adaptive Clarity improves over Clarity Score and other state-of-the-art pre- and post-retrieval prediction approaches.

The work we present is organized as follows: first, in Section 3.2 we will provide an overview of related work. In order to offer some guidance to the levels of accuracy different methods achieve, this section also includes a summary in table form of the correlations found in various publications. Clarity Score is then introduced formally in Section 3.3. Section 3.4 contains an analysis of Clarity Score’s sensitivity to its parameter settings. A similar analysis is also reported for Query Feedback, a post-retrieval method introduced by Zhou and Croft [177]. In Section 3.5, the two proposed changes to the Clarity Score algorithm are outlined. The experimental results, where Clarity Score and our adaptations to it are compared to a number of pre- and post-retrieval predictors are detailed in Section 3.6 and a discussion of the results follows in Section 3.7. The chapter concludes in Section 3.8.

3.2 Related Work

Post-retrieval prediction algorithms can be categorized into different classes, depending on the basic approach they take in order to estimate a query’s result list quality. In this overview, we are going to distinguish the approaches by the type of information they exploit, that is information derived from

- perturbing the query and considering the differences in the respective ranked list of results (Section 3.2.1),
- perturbing the documents of the initially retrieved result list and considering the stability of the ranking (Section 3.2.2),

- perturbing the retrieval approach and considering the diversity of the ranked list of results (Section 3.2.3),
- analysing the ranked list of results of the original query (Section 3.2.4), and lastly,
- Web resources (Section 3.2.5).

Each of the different classes and proposed approaches are outlined below. If not explicitly stated otherwise, the approaches belong to evaluation aspect EA2 of Figure 1.1. Note that some approaches may fall into more than one class. We present those methods in the class of approaches where their most important effect lays.

3.2.1 Query Perturbation

Slightly altering a query and determining the similarity of the result lists of the original and perturbed query indicates the amount of *query drift*, which has been found to be a good indicator of result quality. Query drift refers to the change of focus of a query due to faulty query expansion [107]. A query is considered to be of high quality if slight variations of it do not result in a large change in retrieval result (the query drift is low), while a query whose ranked list of results changes considerably with a small change in the query exhibits a high amount of query drift - a change of focus implying that the originally retrieved ranked list contains a diverse and at least partially non-relevant set of documents.

A direct exploitation of the query drift concept is the *Query Feedback* method, introduced by Zhou and Croft [177]. Query Feedback frames query effectiveness estimation as a communication channel problem. The input is query q , the channel is the retrieval system and the ranked list L is the noisy output of the channel. From the ranked list L , a new, perturbed query q' is generated and a second ranking L' is retrieved with q' as input. The overlap between the lists L and L' is used as query quality score. The lower the overlap between the two rankings, the higher the query drift and thus the lower the predicted effectiveness.

In the *Weighted Information Gain* [177] approach, a number of query variations are derived from a given query and the difference in the probability of occurrence of those variations in the top retrieved documents and in the corpus is used as estimate of retrieval effectiveness. This approach was developed for retrieval models that exploit term dependencies such as the Markov Random Field model [104] and thus the query variations include single term queries, exact phrase queries and unordered window queries, the latter being queries whose constituent terms need to occur in a document within a specified term range. The more the top ranked documents and the corpus as a whole differ with respect to those query variations, the better the estimated quality of the result list. This approach could have also been included in Section 3.2.4, but due to its derivation of different queries we chose to include it here. The experiments reported by Zhou and Croft [177] show that Weighted Information Gain outperforms Query Feedback and that a combination of both in turn overall performs best. The results though are not applicable to our

experiments, as we rely on unigram language models. We leave experimentation on term dependency based retrieval approaches for future work.

Yom-Tov et al. [167] also present an estimator based on query variations. In contrast to Weighted Information Gain though, their query variations are derived from the queries' constituent terms. Those "sub-queries" are then used in retrieval and the result lists of the original query and the sub-queries are compared to each other. The higher the overlap between them, the higher the estimated result list quality. The idea behind this approach is that for well performing queries the result list does not change considerably if only a subset of query terms is used. The two proposed estimators are based on machine learning approaches, and apart from the overlap of the result lists also exploit features such as the retrieval status value of the top ranked document and the number of query terms. The reported experiments suggest that the best predictor performance can be achieved on queries derived from TREC description topics (long queries). It is difficult though to put the results in context, as the predictor baselines in [167] (such as the standard deviation of IDF) are not very strong.

Finally, Vinay et al. [145] propose to perturb the weights that are assigned to each query term, and to consider a result list of high quality, if slight changes of term weights do not lead to a drastically different result list. Among the four approaches they propose this is the weakest performing one. This idea can also be viewed as a generalization of Yom-Tov et al. [167]'s estimators, where each sub-query is formed by setting the weights of all other query terms to zero.

3.2.2 Document Perturbation

The notion of estimating the quality of a result list by its ability to withstand the introduction of noise is based on observations made in retrieval on noisy text corpora. Transforming audio and images to text with automatic speech and optical character recognition leads to corrupted text documents, as the recognition process is imperfect. Experiments on such text corpora have shown that retrieval approaches which are robust in the presence of errors also exhibit a higher retrieval effectiveness on noise free corpora [131]. Translating this observation to query effectiveness prediction leads to the following heuristic: a result list which is stable in the presence of introduced noise is considered to be of high quality, while a result list which is unstable when noise is added to documents (the documents are perturbed) is considered to be of low retrieval effectiveness.

The *Ranking Robustness* approach by Zhou and Croft [176] exploits this heuristic by retrieving a result list for a given query, perturbing the documents by adding or removing terms and then ranking those perturbed documents based on the original query and retrieval approach. The similarity between the original result list and the result list derived from the perturbed documents indicates the robustness. In particular, perturbed are the term frequencies of the query terms occurring in each document by sampling a new frequency value from a Poisson distribution. The similarity between two result lists is determined by Spearman's rank correlation coefficient. Finally this process is repeated a number of times and the average rank correlation

constitutes the robustness score. The higher the score, the better the estimated retrieval effectiveness of the original ranked list of results. A comparison between Clarity Score and Ranking Robustness showed a slightly better performance of the latter. Zhou and Croft [177] also propose a variation of Ranking Robustness, the so-called *First Rank Change* approach, which is modified to be applicable to navigational queries [23]. Instead of comparing the original and perturbed result list, now it is only of interest in how many trials the top ranked document of the original list is also returned at the top of the perturbed result list.

Among the prediction methods proposed by Vinay et al. [145], the best performing one is based on document perturbations. To estimate the retrieval performance of a query, each document in the result list is perturbed by varying degrees of noise and then the perturbed document in turn is used as query. Of interest is the rank, the unperturbed document is retrieved at in response to such a query. When no noise is added to a document, the original (unperturbed) document can be expected to be retrieved at rank one. With increasing levels of noise, the rank of the original document is expected to drop. This rate of change between the level of noise and the drop in rank of the unperturbed document is the measure query quality. The more quickly the rank drops in response to noise, the lower the estimated retrieval performance of a query. The experiments on TREC Vol. 4+5 show the validity of the approach, a Kendall's τ of 0.52 is reported. However, since the retrieval approach in these experiments is TF.IDF and the queries are derived from TREC topic descriptions (instead of TREC topic titles as in most other experiments), it is not possible to directly compare the results to other works.

As opposed to directly altering the terms or term weights of a document as done in the previous two approaches, Diaz [50] proposes a method based on spatial autocorrelation. In this approach a document's retrieval score is replaced by the weighted sum of retrieval scores of its most similar documents in the result list as determined by TF.IDF. The linear correlation coefficient between the original document scores and the perturbed document scores is then used as estimate of result list quality. This method is based on the notion that the result lists of well performing queries are likely to fulfill the cluster hypothesis [143], while poorly performing queries are not. If the cluster hypothesis is fulfilled, we expect the most similar documents to also receive similar retrieval scores by the retrieval system, while in the opposite case, high document similarity is not expressed in similar retrieval scores and the perturbed scores will be very different from the original ones. Note that in [50] this method is referred to simply as $\rho(\tilde{y}, y)$, in our experiments (and in Table 3.1) we denote it with *ACSim* for autocorrelation based on document similarity. The results reported in [50] show that this approach outperforms both Clarity Score and Ranking Robustness on a range of query sets and corpora. An adaptation of this method, where the retrieval scores are modelled by Gaussian random variables is described by Vinay et al. [146].

3.2.3 Retrieval System Perturbation

Different retrieval systems applied to a single corpus return different result lists for a topic, depending on the particular retrieval approach, the approach’s parameter settings, the pre-processing steps applied as well as the corpus content relied upon, such as for instance document content, document titles, anchor text and hyperlink structure. In a controlled setting such as TREC, a very limited number of relevant documents exist per topic (see Table 3.2) and retrieval approaches achieving a high retrieval effectiveness for a topic necessarily have a large amount of overlap among their top retrieved documents. Together with the observation that retrieval systems do not return the same non-relevant documents [134], the following heuristic arises: a topic is easy, that is, it will result in a high retrieval effectiveness, if the document overlap among different retrieval approaches is high. Conversely, a small amount of document overlap indicates a topic whose retrieval effectiveness will be low. In general, approaches in this category evaluate the effectiveness of a topic without considering a particular retrieval system (evaluation aspect EA1 in Figure 1.1) and we speak of *topic* as opposed to *query* since the retrieval approaches may rely on different (TREC) topic parts or different instantiations of the same topic part (different stemming algorithms for instance). The ground truth in this evaluation setup is commonly derived by considering the retrieval effectiveness of each topic across all participating retrieval approaches. The average, median or majority average precision is then utilized as ground truth effectiveness.

The first approach in this direction is *AnchorMap*, proposed by Buckley [25]. In his work, the document overlap between two rankings is equivalent to the mean average precision a ranking achieves if the documents of a second ranking are considered to be the relevant ones (the documents are “anchored”). The reported correlation with the ground truth is significant at $r = 0.61$. However, due to the small scale of the study – 30 topics and 8 retrieval systems – it is not possible to draw further conclusions from the result.

Counting the number of unique documents among the result lists of different retrieval approaches was suggested by Takaku et al. [136]. The larger the count, the more diverse the result lists and thus the more difficult the topic is estimated to be. The reported correlation of $r = -0.34$ suggests that there is some relationship between topic difficulty and the number of unique documents. However, due to the size of the experiment (14 retrieval systems) and the type of topics (navigational topics from NTCIR) it remains unclear if these results will hold for informational queries.

Aslam and Pavlu [7] propose to determine the document overlap between retrieval systems by the Jensen-Shannon divergence [100] of their result lists. The authors experiment with a wide range of TREC data sets and report consistently high correlations, which indicates that the topic difficulty inherent to a collection is easier to estimate than the query effectiveness for a particular retrieval approach.

3.2.4 Result List Analysis

The ranked list of results can either be evaluated with respect to the corpus or by comparing the documents in the result list to each other without any further frame of reference. Comparing the retrieved result list to the collection as a whole is a measure of the result list's *ambiguity*. If the top retrieved results appear similar to the collection, the query is estimated to be difficult, as the results are not distinct from the entire corpus which covers many topics. On the other hand, if the top retrieved results are homogeneous and different from the corpus, the query's result list is estimated to be of high quality.

Clarity Score [45], which will be covered in more detail in Section 3.3, is based on the intuition that the top ranked results of an unambiguous query will be topically cohesive and terms particular to the topic will appear with high frequency. The term distribution of the top ranked results of such a query will be different from the general term distribution, which is derived from the entire corpus of documents. In contrast, a retrieval system will return results belonging to number of different topics when the query is ambiguous, and thus the term distribution will be less distinguishable from the corpus distribution. The higher the Clarity Score, the more distinct the top ranked results are from the corpus. In this case, the results are estimated to be unambiguous and therefore the estimated quality of the query is high. While Clarity Score is based on the Language Modeling framework, the same idea of comparing the term distribution in the top retrieved documents to the corpus as a query quality measure has been introduced for the Divergence From Randomness retrieval model [5] by Amati et al. [4]. In the work of Diaz and Jones [52], it is proposed to linearly combine Clarity Score with temporal features derived from the top ranked results. In particular in news corpora, distinctive temporal profiles exist for certain terms, such as *Christmas* which will occur mostly in articles published around December of each year, while a term such as *explosion* is likely to occur in bursts across the year depending on current events. This combination of article content and article publication time based query quality measures proved to lead to considerably higher correlations with average precision than Clarity Score alone. The two corpora used in the experiments were derived from TREC collections to only include newspaper articles, making them particularly suitable for the task. The same approach is unlikely to lead to improvements on WT10g and GOV2 though.

Clarity Score has also been applied to tasks such as selective query expansion [46] and the automatic identification of extraneous terms in long queries [94].

Carmel et al. [30] hypothesize that query difficulty is positively correlated with the distances between the query, the corpus and the set of relevant documents. The motivational experiments show that as expected the distance between the set of relevant documents and the set of all documents (the corpus) exhibits a positive correlation with retrieval effectiveness, an observation that is similar to the motivation for Clarity Score. Since in the topic difficulty setting the set of relevant documents is unknown, it is proposed to approximate this set by performing an initial retrieval and then selecting those documents from the result list that lead to the smallest distance to the query. The distances derived from the query, corpus and the approxi-

mation of the set of relevant documents are then used as features in a support vector machine [44, 144]. The approach is evaluated on GOV2 and queries derived from title topics 701-800. The reported linear correlation coefficient ($r = 0.36$) though lacks behind other reported approaches.

The *Clustering Tendency* method developed by Vinay et al. [145] deems a result list to be of high quality if the documents are tightly clustered, whereas the lack of clusters in the result list indicates poor retrieval effectiveness. In contrast to the previously described approaches, this method does not compare the result list to the corpus, instead, the amount of clustering is derived from the top retrieved documents alone. In this approach, documents are points in a high-dimensional space (vector space model). Then, random points are generated and two distances are recorded for each generated sample: the distance between the random point and the nearest document point d_D and the distance between d_D and its nearest document point neighbor. A large difference in those two distances indicates a high quality result list: the documents are highly clustered as their distances to each other are lower than their distances to random points. The advantage of such an approach is that we do not require collection statistics, on the other hand this might also make us miss vital information. A corpus that contains sports documents only, and whose top 100 results are about sports are not really tightly clustered, whereas they appear clustered when we deal with a general news corpus. The reported results show a good estimation performance ($\tau = 0.44$). It should be noted though, that the distances between documents are determined based on query terms only, which works well for longer queries (in the experiments TREC description topics were used), the effect on short queries is not known.

Instead of considering the content of the top retrieved documents, recent studies have also investigated the possibility of deriving quality measures directly based on the retrieval scores of the top ranked documents. The most basic possibility is to estimate the quality of a result list by the retrieval score assigned to the top retrieved document as proposed by Tomlinson [139]: the higher the retrieval score, the better the result list is estimated to be. The performance of this approach is naturally dependent on the retrieval approach (which in turn determines what retrieval scores to assign to each document) and the query set. Depending on the retrieval model settings, the results reported in [139] for the Hummingbird SearchServer vary between $\tau = 0.23$ and $\tau = 0.35$ for queries derived from TREC title topics and between $\tau = 0.26$ and $\tau = 0.43$ for queries derived from TREC description topics. The observation that this approach is better suited for longer queries was also confirmed by Yom-Tov et al. [167] who evaluated the Juru search engine.

Shtok et al. [130] and Perez-Iglesias and Araujo [118] experiment with estimating the coverage of query aspects in the ranked list of results by deriving the retrieval scores' standard deviation, possibly normalized by a query dependent corpus statistic. It is hypothesized, that a high standard deviation indicates a high "query-commitment" [130] and the absence of aspects unrelated to the query. This indicates a result list of high quality. Conversely, if the retrieval scores of the top ranked documents exhibit a low score diversity, the result list is estimated to be dominated by aspects unrelated to the query and it is therefore considered to be of

poor quality. The work by Lang et al. [96] is also similar in spirit. Here, each query term is considered as a separate concept and the more concepts are covered in the result list, the better the estimated result quality. The *coverage score* of a query is defined as the sum of the term coverage scores weighted by the terms' importance. The term coverage scores in turn are estimated based on the retrieval scores of the top ranked documents. The advantage of retrieval score based approaches is the very low complexity, compared to approaches relying on document content, or document and query perturbations. At the same time the reliance on retrieval scores can also be considered a drawback, as such approaches require collaborating search systems that make the retrieval status values available. The results reported in [96, 130] show the potential of these approaches: retrieval score based methods achieve similar or higher correlations than the evaluated document content based approaches (including Clarity Score).

3.2.5 Web Resources

A number of recent studies take advantages of resources from Web search engines such as interaction logs, query logs and the Web graph (\mathcal{W}). We describe these approaches here as they offer valuable insights. However, we admit that we cannot apply any of the insights directly to our own work due to the unavailability of such resources to us. A second type of studies we describe in this section relies on freely available Web resources such as the Open Directory Project¹ (ODP) to infer the quality of search results.

Jensen et al. [80] infer the difficulty of a query on the Web by submitting it to different search engines, collecting the presented snippets of the top retrieved results and extracting thirty-one “visual clues” from these snippets such as the percentage of character n-grams of the query appearing in the snippet, the snippet title and the URL. An SVM regression approach is then applied to train a model. The reported results confirm the validity of the approach, the Spearman rank correlation between average precision at 10 documents (averaged over all search engines) and the quality estimate is $\rho = 0.57$. Since all baseline approaches are pre-retrieval predictors and this approach amounts to a post-retrieval approach it remains to be seen how its performance will compare against other approaches relying on the result list.

Leskovec et al. [99] propose the utilization of search result dependent subgraphs of \mathcal{W} (on the URL and domain level) to train search result quality classifiers. The nodes of the top ranked Web pages retrieved in response to a query are located in \mathcal{W} ; together with their connecting edges they form the *query projection graph*. A second subgraph, the *query connection graph* is generated by adding nodes and edges to the query projection graph until all nodes of the top ranked results are in a single connected component. A total of fifty-five features, most of them topological in nature, are derived from these two subgraphs, the query and the search result list. Together they are used in a Bayesian network classifier. The approach, evaluated on nearly 30000 queries, was found to achieve a high degree of classification accuracy.

¹<http://www.dmoz.org/>

The drawback of such a reliance on the Web graph is the complexity of the approach, in particular in finding the query connection graph.

A step further goes the research reported by White et al. [157], where “supported search engine switching” is investigated. Here, the aim is to predict which search engine of a given set of engines is going to provide the best results for a particular query. A large interaction log, containing search sessions of more than five million users across Google, Yahoo! and MS Live, was analyzed to find a set of useful features for the task of determining which of two rankings is of higher quality. Three types of features, used in a neural network based classifier, were found to lead to a high prediction accuracy: features derived from the two result rankings, features based on the query and features based on the similarity between query and result ranking. The evaluation on a data set of 17000 queries showed that automatic engine switching can considerably improve result precision.

Predicting when to switch between states is also explored by Teevan et al. [137], whose goal is to predict whether to switch query personalization on or off. The motivation for this work stems from the observation, that not all queries perform equally well when user dependent factors, such as search history and user profile, are taken into account. While ambiguous queries benefit from personalization, the result quality of non-ambiguous queries can deteriorate. In order to predict query ambiguity, the authors rely on a query log of more than 1.5 million users and 44000 distinct queries. A Bayesian dependency network is learned with forty features extracted from the query, the result list and the query log, including the average number of results, the click position, the time a query is issued and the amount of seconds passed until a result is clicked. Click entropy, which is the variability in the clicked results across users, is one of the evaluated query ambiguity measures. The results show that the trained model predicts the click entropy based ambiguity with high accuracy. Notably, Clarity Score is also listed as one of the features. Its correlation with click entropy is reported as approximately zero, similarly to most other features that are not based on query log information. The reason of this result remains unclear, though one possible explanation is that Clarity Score determined on the title and summary of the top 20 retrieved results (as done here) is not as effective as on the full document content of possibly hundreds of documents. Furthermore, the ranking produced by a Web search engine is derived from a number of sources of evidence, instead of document content only, which might also influence the performance of Clarity Score.

Finally, Collins-Thompson and Bennett [40] and Qiu et al. [123] propose to exploit the ODP to measure a query’s ambiguity. Although assigning precomputed ODP categories to each document [40] and relying on them to calculate Clarity Score and related approaches has the advantage of a low computational overhead, the reported results are not convincing: the maximum correlations achieved are $\tau = 0.09$ and $\tau = 0.13$ on the WT10g and GOV2 corpora respectively. The ODP also provides a search service, which given a query as input, returns not only a list of result pages but also a list of ODP categories. In [123], an ODP category list is determined for each term of a query and the overlap between the different lists of a query is used as an indicator of query ambiguity. A problem of the approach is that it

cannot assign ambiguity scores to queries consisting of a single term only. The high correlations reported (up to $\rho = 0.81$) were achieved on queries derived from the topics of the TREC 2003/04 Novelty track. In this track, only 25 documents exist per topic and thus the entire document collection consists of merely 2500 documents. It is not known how the reported results will translate to larger corpus sizes.

3.2.6 Literature Based Result Overview

While outlining the various approaches in the previous sections, we have largely refrained from comparing the effectiveness of different algorithms. The reason for this is the diversity of the test corpora, retrieval approaches, topic and query sets and evaluation measures, that have been employed to evaluate these algorithms. It is only possible to draw conclusions from evaluations performed on exactly the same setup. As was shown in Chapter 2, a small change in the parameter settings of a retrieval approach can already influence the correlation a query performance prediction method achieves. This observation also holds for post-retrieval methods as will become clear in the result section of this chapter.

Due to the complexity of most approaches, it is not possible to perform an evaluation across all methods. In order though to give some indication of the success of selected methods, we provide an overview of the correlations they exhibit on TREC and NTCIR data sets in Table 3.1. All correlations are taken from the cited publications. If a publication contains several proposed prediction methods, we include the best performing ones.

For each method included in Table 3.1, the overview contains the evaluation aspect that is investigated, the effectiveness measure relied upon as ground truth, the topic set and where applicable the part of the TREC/NTCIR topic the queries are derived from, either title (T), description (D), any part including the narrative (-) or unknown (?). The last three columns list the correlation coefficients.

The respective evaluation aspect – EA1 or EA2 – determines how the ground truth is derived. The ground truth of methods that predict the effectiveness of a set of queries for a particular retrieval approach (EA2) is most often the average precision (AP), some results also exist for precision at 10 documents (P@10) and reciprocal rank (RR). The latter measure is used in instances where navigational query sets are evaluated. The column *Models/#Runs* lists the retrieval approach the ground truth effectiveness is derived from: either Language Modeling (LM), TFIDF, BM25, the Markov Random Field model (MRF) or the Divergence From Randomness model (DFR).

The ground truth of methods that aim to predict the topic difficulty inherent to the corpus (EA1) is derived from a set of retrieval approaches. When diverse retrieval approaches achieve a low retrieval effectiveness a topic is deemed difficult for a corpus. In this setting, the median, the average or the majority AP value across the retrieval runs participating in the experiment form the ground truth. The column *Models/#Runs* specifically indicates how many TREC or NTCIR runs are relied upon.

Table 3.1: Overview of selected post-retrieval query and topic effectiveness predictors. **Models/#Runs** describes the number of runs or the retrieval model used, depending on the evaluation aspect **Eval. Asp.**: EA1 (How difficult is a topic in general?) and EA2 (How difficult is a topic for a particular system?). **T-N** indicates which topic part is used: title (T), description (D), any part including the narrative (-) or unknown (?). **Evaluation Measure** lists the effectiveness measure relied upon as ground truth: average precision (AP), precision at 10 documents (P@10), reciprocal rank (RR) and the average and median AP value over all evaluated runs (TREC av. AP and TREC med. AP respectively). The correlations reported in each publication are shown in the last three columns: the linear correlation coefficient r , Kendall's τ and Spearman's ρ .

	Topics	Models/#Runs	T-N	Eval. Asp.	Evaluation Measure	r	τ	ρ
Clarity Score[45]: divergence between the query language model and the collection language model	201-250	LM	D	EA2	AP			0.490
	251-300	LM	T	EA2	AP			0.459
	351-400	LM	T	EA2	AP			0.577
	401-450	LM	T	EA2	AP			0.494
	351-450	LM	T	EA2	AP			0.536
<i>Info_{B_{o2}}</i> [4]: divergence between query term frequencies of the collection and result list	100 topics (TREC Robust 2003)	DFR	D	EA2	AP	0.52		
	\approx 100 topics from AP corpus	LM	?	EA2	AP	0.52		
Temporal Clarity [52]: linear regression with Clarity Score and two features of temporal profiles based on the creation dates of the top retrieved documents	\approx 100 topics from WSJ corpus	LM	?	EA2	AP	0.60		
	30 topics from 301-450	8 auto. TREC runs	D	EA1	majority AP	0.608		
AnchorMap[25]: document overlap between ranked lists measured by mean average precision	301-450,601-700	Hummingbird	T	EA2	AP		0.35	
	301-450,601-700	Hummingbird	D	EA2	AP		0.43	
Top Score[139]: retrieval score of the top retrieved document	301-450,601-650	Juru	T	EA2	AP		0.305	
	301-450,601-650	Juru	T	EA2	P@10		0.268	
	451-550	Juru	D	EA2	AP		0.202	
	451-550	Juru	T	EA2	P@10		0.175	
Histogram[167]: document overlap between the result lists of the full query and its sub-queries; feature histogram based machine learning approach	301-450,601-650	Juru	D	EA2	AP		0.439	
	301-450,601-650	Juru	D	EA2	P@10		0.360	
	451-550	Juru	T	EA2	AP		0.143	
	451-550	Juru	T	EA2	P@10		0.187	
Pool size[136]: number of unique documents in the pool of top 100 retrieved documents	269 NP NTCIR-5 topics	14 NTCIR runs	T	EA1	av. RR	-0.342		
	301-450,601-650	TFIDF	D	EA2	AP		0.441	
Document clustering tendency[145]	301-450,601-650	TFIDF	D	EA2	AP		0.521	

	Topics	Models/#Runs	T-N	Eval. Asp.	Evaluation Measure	Correlations		
						r	τ	ρ
Combination of topic distances[30]: SVM based machine learning approach	701-800	Juru	T	EA2	AP	0.362		
	201-250	LM	D	EA2	AP	0.613	0.548	
	251-300	LM	T	EA2	AP	0.454	0.328	
	301-450,601-700	LM	T	EA2	AP	0.550	0.392	
	701-750	LM	T	EA2	AP	0.341	0.213	
	751-800	LM	T	EA2	AP	0.301	0.208	
	251-300	all TREC runs	-	EA1	TREC av. AP	0.623	0.469	
	301-350	all TREC runs	-	EA1	TREC av. AP	0.698	0.491	
	351-400	all TREC runs	-	EA1	TREC av. AP	0.722	0.623	
	401-450	all TREC runs	-	EA1	TREC av. AP	0.770	0.615	
JS divergence[7]: diversity between the result lists of multiple retrieval systems	301-450,601-700	all TREC runs	-	EA1	TREC av. AP	0.695	0.530	
	701-750	all TREC runs	-	EA1	TREC av. AP	0.682	0.502	
	751-800	all TREC runs	-	EA1	TREC av. AP	0.581	0.440	
	301-450,601-700	MRF	T	EA2	AP	0.468		
	701-800	MRF	T	EA2	AP	0.574		
	801-850	MRF	T	EA2	AP	0.464		
	252 NP-05 topics	MRF	-	EA2	RR	0.458		
	181 NP-06 topics	MRF	-	EA2	RR	0.478		
	301-450,601-700	MRF	T	EA2	AP	0.464		
	701-800	MRF	T	EA2	AP	0.480		
Query Feedback[177]: overlap between the original result list ℓ_O and the result list derived by constructing a query from ℓ_O	801-850	MRF	T	EA2	AP	0.422		
	252 NP-05 topics	MRF	-	EA2	RR	0.440		
	181 NP-06 topics	MRF	-	EA2	RR	0.386		
	252 NP-05 topics	MRF	-	EA2	RR	0.525		
	181 NP-06 topics	MRF	-	EA2	RR	0.515		
	701-800	MRF	T	EA2	AP	0.637		
	801-850	MRF	T	EA2	AP	0.511		
	100 topics (TREC Novelty 2003/04)	LM	T	EA2	AP			0.597
	70 topics (TREC Novelty 2003/04)	LM	T	EA2	AP			0.808

	Topics	Models/#Runs	T-N	Eval. Asp.	Evaluation Measure	Correlations		
						r	τ	ρ
ACSim ($\rho(y, \hat{y})$) [50]: spatial autocorrelation based on document similarity	201-250	LM	D	EA2	AP	0.650	0.513	
	251-300	LM	T	EA2	AP	0.486	0.357	
	301-450,601-700	LM	T	EA2	AP	0.527	0.373	
	701-750	LM	T	EA2	AP	0.540	0.454	
	751-800	LM	T	EA2	AP	0.439	0.383	
	201-250	LM	D	EA2	AP		0.615	
Covering Topic Score (IM2/CD2) [96]: coverage of topic concepts in the result list	251-300	LM	T	EA2	AP		0.422	
	301-450, 601-700	LM	T	EA2	AP		0.454	
	701-750	LM	T	EA2	AP		0.313	
	751-800	LM	T	EA2	AP		0.356	
	301-450,601-650	BM25	T	EA2	AP		0.328	
	301-450,601-650	BM25	D	EA2	AP		0.345	
AP Scoring [146]	451-550	LM	T	EA2	AP		0.091	
	701-850	LM	T	EA2	AP		0.130	
	301-450,601-700	BM25	T	EA2	AP	0.55	0.41	
	201-250	LM	D	EA2	AP	0.556	0.414	
	251-300	LM	T	EA2	AP	0.431	0.300	
	301-450,601-700	LM	T	EA2	AP	0.563	0.419	
Topic Prediction $\frac{\Delta_{QR}}{\Delta_{QG}}$ [40]	451-550	LM	T	EA2	AP	0.527	0.303	
	Ranking Dispersion [118]							
	Normalized Query-Commitment (NQC) [130]: standard deviation of retrieval scores							

3.3 Clarity Score

In this section, the Clarity Score query effectiveness estimator will be explained in more detail. To compute Clarity Score, the ranked list of documents returned for a given query is used to create a query language model [97] where terms that often co-occur in documents with query terms receive higher probabilities:

$$P_{qm}(w) = \sum_{D \in R} P(w|D)P(D|Q). \quad (3.1)$$

R is the set of retrieved documents, w is a term in the vocabulary, D is a document, and Q is a query. In the query model, $P(D|Q)$ is estimated using Bayesian inversion:

$$P(D|Q) = P(Q|D)P(D) \quad (3.2)$$

where the prior probability of a document $P(D)$ is zero for documents containing no query terms.

Typically, the probability estimations are smoothed to give non-zero probability to terms not appearing the query, by redistributing some of the collection probability mass:

$$\begin{aligned} P(D|Q) &= P(Q|D)P(D) \\ &= P(D) \prod_i P(q_i|D) \\ &\approx P(D) \prod_i \lambda P(q_i|D) + (1 - \lambda)P(q_i|C) \end{aligned} \quad (3.3)$$

where $P(q_i|C)$ is the probability of the i th term in the query, given the collection, and λ is a smoothing parameter. The parameter λ is constant for all query terms, and is typically determined empirically on a separate test collection.

Clarity Score is the Kullback-Leibler (KL) divergence between the query language model P_{qm} and the collection language model P_{coll} :

$$D_{KL}(P_{qm}||P_{coll}) = \sum_{w \in V} P_{qm}(w) \log \frac{P_{qm}(w)}{P_{coll}(w)}. \quad (3.4)$$

The larger the KL divergence, the more distinct is the query language model from the collection language model. If the documents of the ranked list are very similar to each other, Clarity Score assigns a relatively high score, as the divergence between the query language model and the collection language model will be large. Ambiguous queries on the other hand are hypothesized to result in ranked lists that are not topically homogeneous, which leads to a lower divergence between the two language models.

3.3.1 Example Distributions of Clarity Score

In this section, we experimentally assess whether the homogeneity assumption described above holds. For each query of our query sets, we calculate the Clarity Scores of three ranked lists, namely the lists of:

- the x relevant documents,
- a random sample of x documents from the pool of non-relevant documents, and,
- a random sample of x documents from the pool of all documents in the collection containing at least one title topic term.

To derive the ranked lists for the relevant and the pool of non-relevant documents, we rely on the relevance judgments (the so-called *qrrels*), available for each TREC test collection. For each topic, the *qrrels* contain the relevant as well as the judged non-relevant documents. The number x of documents to use is topic dependent and equal to the total number of relevant documents available for a topic. Table 3.2 shows the minimum, average and maximum number of relevant documents x across all topics of a corpus. While in all test corpora topics occur with very few relevant documents, the average number of relevant documents for the GOV2 collection is significantly higher than for the topics of TREC Vol. 4+5 and WT10g.

We distinguish between two random samples: the *non-relevant* random sample and the *collection-wide* random sample. The non-relevant random sample is derived from documents judged as non-relevant in the *qrrels*. As TREC assessments are made of a pool of documents which have been returned as the top ranked documents by participating systems, it can be expected that those non-relevant documents are somewhat close in spirit to the relevant documents from the point of view of keyword based retrieval. As the number of judged non-relevant documents is always larger than the number of relevant documents, x samples are drawn. This sampling process is repeated five times and the average Clarity Score of those five iterations is reported. Note that only documents containing at least 50 terms were considered. For the collection-wide sample, we rely on our stopped and stemmed indices. All documents in the collection, that contain at least one of the title topic terms and have a length of 50 or more terms (including stopwords), are used as sample space and as before sampling is performed five times and the average is reported.

Corpus	Topics	#Relevant Documents		
		Min.	Average	Max.
TREC Vol. 4+5	301-450	3	93	474
WT10g	451-550	1	60	519
GOV2	701-850	4	181	617

Table 3.2: Minimum, average and maximum number of relevant documents in the relevance judgments.

If the homogeneity assumption holds, we can expect a noticeable difference between the *relevant*, *non-relevant* and *collection-wide* Clarity Scores. Ideally, we would expect the scores of the *relevant* lists to lie in a narrow band, as no non-relevant document enters the language model and the ranked lists of results are unambiguous. The Clarity Scores of the *non-relevant* lists are expected to be somewhat lower, but still higher than those of the *collection-wide* lists as the non-relevant documents were

mistaken to be relevant by at least one retrieval system, whereas the *collection-wide* lists are generally created from a very large pool of random documents. In cases, where the title topic consists of a single very specific term, the pool of random documents will become very small and no large difference in scores can be expected. This effect is observed rarely though.

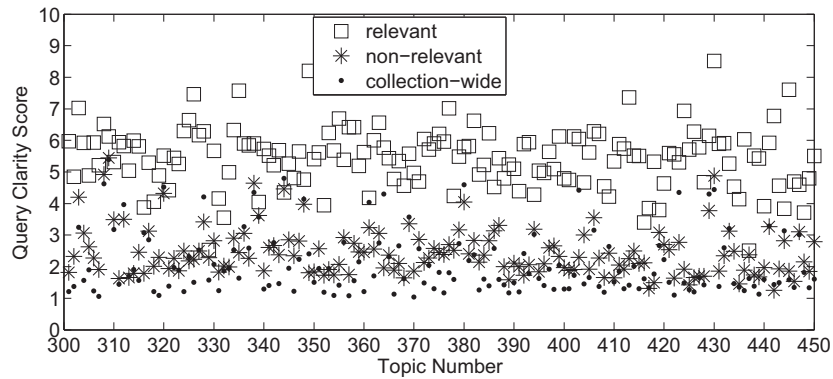
The results of this experiment are shown in the form of scatter plots in Figure 3.1. Each point marks the Clarity Score of a query for a particular type of ranked list, either *relevant*, *non-relevant* or *collection-wide*. In general, the results are as expected, that is the Clarity Scores of the lists of relevant documents are higher than those of the non-relevant and the collection-wide ones. However, there are differences visible in the quality of separation between the three plots. Figure 3.1c contains the results of the GOV2 collection. Here, in all instances, the scores of the lists of relevant documents are higher than those of the two random samples and furthermore, there are only 22 cases (out of 150) where the collection-wide samples achieve a higher score than the non-relevant samples. Slightly less well separated are the queries of TREC Volumes 4+5 (Figure 3.1a); for one query², the list of relevant documents has a slightly lower score than the two random samples and in 33 additional cases the collection-wide samples are considered more homogeneous than the non-relevant samples. The results of the WT10g collection in Figure 3.1b are worst with respect to the separability of the different list types. There exist nine queries where the Clarity Scores of the lists of relevant documents are lower than the scores of one or both random samples. For a further 33 queries, the scores of the collection-wide samples are higher than the non-relevant random samples. These results indicate, that Clarity Score is likely to perform better on GOV2 and TREC Vol. 4+5 than on WT10g, since the separation between the relevant and random samples is considerably clearer for them.

3.4 Sensitivity Analysis

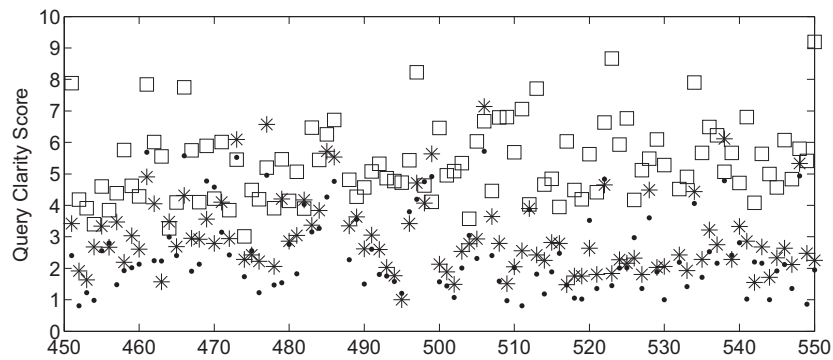
During the analysis of Clarity Score’s homogeneity we assumed the number of feedback documents x to be query dependent, equaling the number of relevant documents existing for a topic. This knowledge over the number of relevant documents is, of course, not available in practical applications and the original Clarity Score algorithm utilizes a uniform setting of x across all queries, with $x = 500$ being reported to be a good value [45].

In this section, we investigate the influence of different factors affecting effectiveness prediction quality by giving examples of the behavior of Clarity Score and Query Feedback [177] as their (i) parameters, (ii) the retrieval setting, (iii) the collections and (iv) the query sets vary. In particular we are interested how sensitive Clarity Score actually is to the setting of x and how well it can perform for the WT10g collection, having in mind the homogeneity analysis of Section 3.3.1. Additionally, we perform a similar analysis for the Query Feedback algorithm, as it

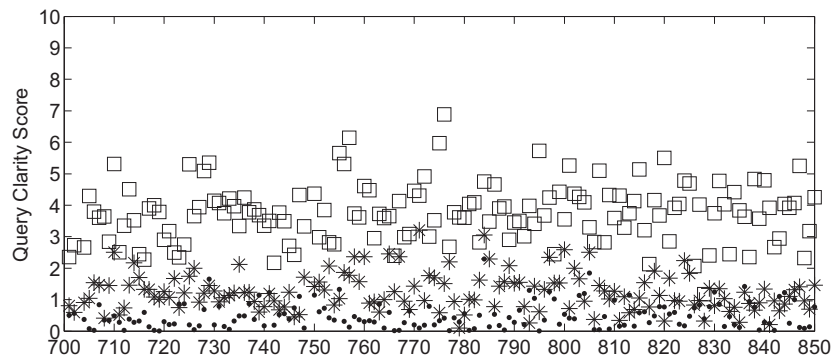
²Title topic 344: “Abuses of E-Mail”; the corresponding stemmed and stopword-free query is “abuse e mail”



(a) Queries 301-450 (TREC Vol. 4+5)



(b) Queries 451-550 (WT10g)



(c) Queries 701-850 (GOV2)

Figure 3.1: Distribution of Clarity Scores of the lists of relevant documents, sampled lists of non-relevant documents and sampled lists of collection-wide documents.

has been shown to achieve a good prediction performance across various TREC test collections. The parameters of Query Feedback are the number $t = |\mathbf{q}'|$ of terms \mathbf{q}' consists of and the number of top documents s for which the overlap between the two rankings L and L' is considered.

In Chapter 2 we already observed that the retrieval approach relied upon has a considerable influence on the accuracy of prediction algorithms. As will become evident shortly, the same observation holds for post-retrieval algorithms. For this reason, we evaluate Clarity Score and Query Feedback for a number of param-

ters of the Language Modeling with Dirichlet smoothing approach, in particular $\mu = \{100, 500, 1000, 1500, 2000, 2500\}$. As in all experiments, we derive the queries from the TREC title topics.

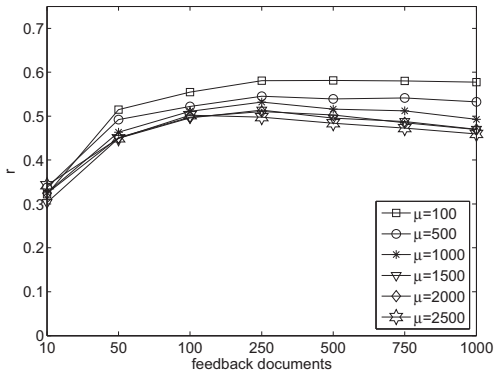
3.4.1 Sensitivity of Clarity Score

Figure 3.2 shows the development of Clarity Score’s performance in terms of the linear correlation coefficient (the trends are similar for Kendall’s τ). The number x of feedback documents is evaluated for the range of $x = \{10, 50, 100, 250, 500, 750, 1000\}$. Figures 3.2a, 3.2b and 3.2c display the behavior of the three different query sets of TREC Vol. 4+5. While queries 301-350 are relatively insensitive to the specific number of feedback documents and do not show much change in performance once 250 feedback documents are reached, queries 351-400 exhibit a very different behavior. At 10 feedback documents and $\mu = 2000$, the linear correlation coefficient is as high as $r = 0.66$, while at 1000 feedback documents the correlation has degraded to $r = 0.27$. Finally, queries 401-450 show a continuous increase in r for the lower levels of smoothing, while for $\mu = 1500$ and above, Clarity Score’s performance peaks at 250 feedback documents. In more general terms, for this collection the observation holds that low levels of smoothing favor a good Clarity Score performance: across most settings of x , the lowest level of smoothing ($\mu = 100$) leads to the highest correlation.

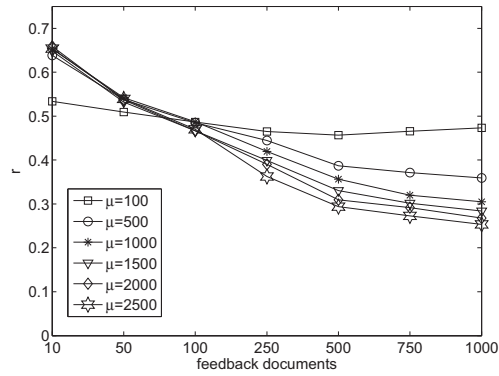
In Section 3.3.1, we hypothesized that Clarity Score’s performance will be worst for WT10g, due to the insufficient separation between the scores of the relevant, the non-relevant and the collection-wide randomly drawn documents. This hypothesis is now empirically confirmed when considering Figures 3.2d and 3.2e. They contain the results of the two query sets of the WT10g collection. Compared to the other query sets, the prediction performance is considerably lower, achieving in the most favorable setting $r = 0.43$. The influence of the level of smoothing is visible, but less clear: while for queries 451-500 $\mu = 100$ gives the highest correlation, the same level of smoothing leads to a low performance when considering queries 501-550. The influence of the number of feedback documents also varies; for queries 451-500, at $x = 10$ the Clarity Score’s performance peaks for all but one smoothing level ($\mu = 100$). In contrast, for queries 501-550 the highest performance is achieved when x is set to between 250 and 1000, depending on μ .

Finally, the results for the query sets of the GOV2 corpus are shown in Figures 3.2f, 3.2g and 3.2h. Here, overall a greater amount of smoothing leads to a better performance and the optimal setting of x varies between 100 and 250.

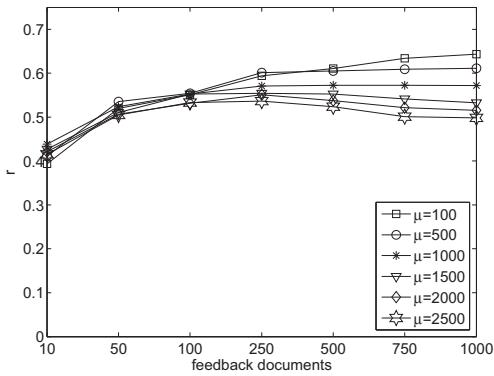
The setting of μ that leads to the highest correlation, often does not result in the best retrieval performance as measured in MAP. Consider Table 2.2, which contains the overview of the retrieval effectiveness for all query sets and various settings of μ . In all but one case $\mu \geq 1000$ results in the highest retrieval effectiveness. However, for queries 451-500 for instance, the highest linear correlation coefficient ($r = 0.43$) is achieved for the setting of $\mu = 100$ and $x = 500$ feedback documents. The MAP of this retrieval run is only 0.15 though. This is significantly worse than the MAP of the best performing run (0.21), which in turn leads to a maximum predictor correlation



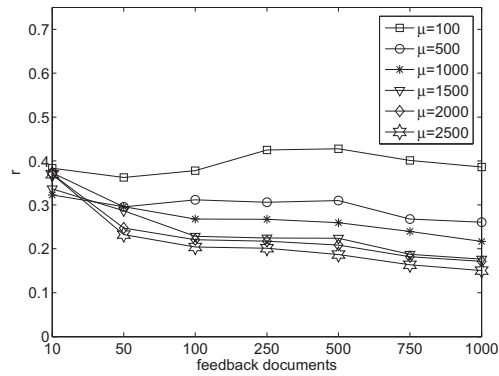
(a) Queries 301-350



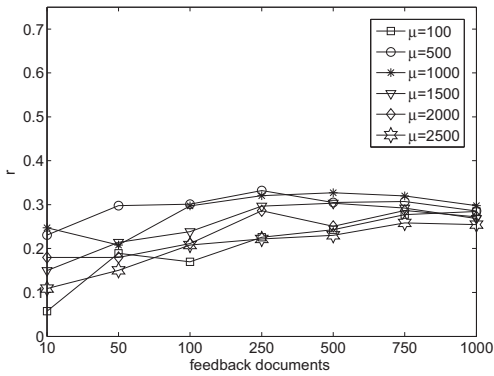
(b) Queries 351-400



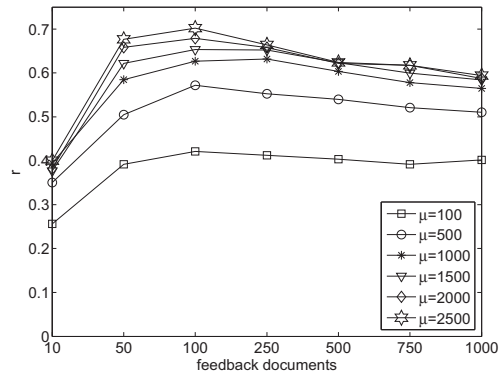
(c) Queries 401-450



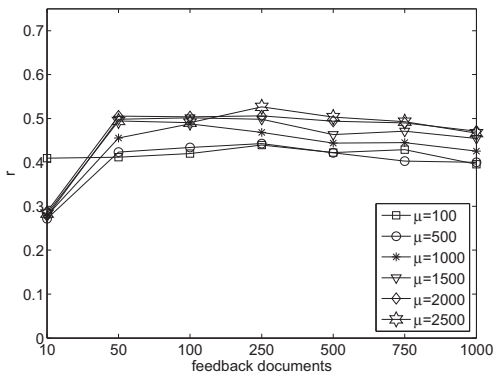
(d) Queries 451-500



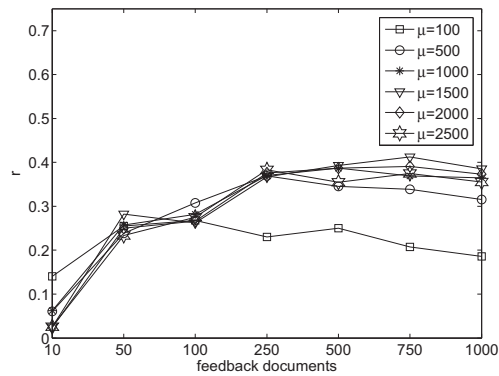
(e) Queries 501-550



(f) Queries 701-750



(g) Queries 751-800



(h) Queries 801-850

Figure 3.2: Sensitivity of Clarity Score towards the corpus, the smoothing parameter μ and the number of feedback documents.

Corpus	Queries	Best	Standard	Worst
TREC Vol. 4+5	301-350	0.545	0.539	0.338
	351-400	0.659	0.311	0.268
	401-450	0.573	0.573	0.438
WT10g	451-500	0.323	0.260	0.217
	501-550	0.287	0.251	0.180
GOV2	701-750	0.635	0.603	0.406
	751-800	0.488	0.444	0.279
	801-850	0.387	0.387	0.062

Table 3.3: Linear correlation coefficient r of the best, standard (500 feedback documents) and worst performing Clarity Score with respect to the retrieval run with the highest retrieval effectiveness as given in Table 2.2.

Corpus	Queries	Best	Standard	Worst
TREC Vol. 4+5	301-350	0.436	0.420	0.302
	351-400	0.503	0.217	0.155
	401-450	0.367	0.305	0.305
WT10g	451-500	0.300	0.129	0.118
	501-550	0.243	0.223	0.053
GOV2	701-750	0.475	0.415	0.257
	751-800	0.377	0.330	0.243
	801-850†	0.247	0.235	0.061

Table 3.4: Kendall’s τ of the best, standard (500 feedback documents) and worst performing Clarity Score with respect to the retrieval run with the highest retrieval effectiveness as given in Table 2.2.

of $r = 0.32$.

To stress the point that the standard setting of 500 feedback documents may not always be adequate for Clarity Score, we present in Tables 3.3 and 3.4 the linear correlation coefficient r and Kendall’s τ that Clarity Score achieves with 500 feedback documents (standard) as well as the correlations of the best and worst performing feedback document setting. It is evident, that the feedback parameter is important for the accuracy of the Clarity Score algorithm and a wrong setting of this parameter can lead to poor results.

3.4.2 Sensitivity of Query Feedback

Figure 3.3 shows Query Feedback’s sensitivity to changes in its parameter settings exemplary for queries 351-400 (reported as linear correlation coefficient r) and queries 451-500 (reported as Kendall’s τ for comparison). The parameters s and t are evaluated for the range of $s = \{20, 50, 100\}$ and $t = \{2, 5, 10, 20\}$ respectively. Noticeable, as for Clarity Score, is the dependency on the correct parameter setting. The correlations achieved fluctuate widely, depending on s and t but also depending on the query set. For instance, for queries 351-400, the setting of $s = 20$ results

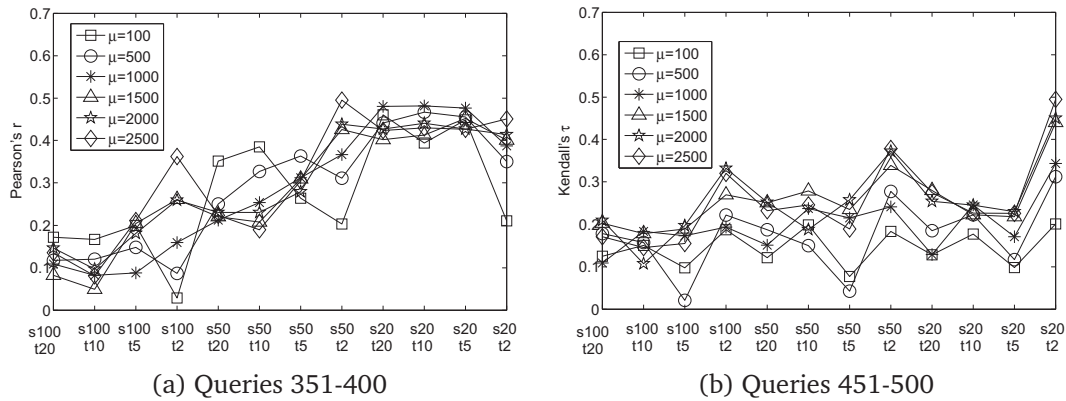


Figure 3.3: Sensitivity of Query Feedback towards its parameters and the smoothing parameter μ of the retrieval approach (language modeling with Dirichlet smoothing).

in a very stable and good performance across all t except for $t = 2$, whereas for queries 451-500, $t = 2$ performs best across all settings of s . Finally, the effect of the level of smoothing on the algorithm's quality is generally reversed compared to Clarity Score: the lower the level of smoothing μ , the less well the Query Feedback algorithm performs.

The conclusion to be drawn from this analysis is that both Clarity Score and Query Feedback can be very sensitive to both the initial retrieval parameter tuning, as well as their own parameters. Furthermore, parameters tuned to one query set do not produce reliable results for other query sets. Even when the query set and the collection are fixed, the performance of the predictors vary depending on the parameter settings of the retrieval approach.

3.5 Clarity Score Adaptations

In this section we introduce our proposed adaptations to Clarity Score. First the approach to setting the number of feedback documents automatically is described, followed by the frequency dependent term selection.

3.5.1 Setting the Number of Feedback Documents Automatically

In the literature, setting the number of feedback documents to a fixed value for all queries is the standard approach. Cronen-Townsend et al. [45] suggest that the exact number of feedback documents used is of no particular importance and 500 feedback documents are proposed to be sufficient. In Section 3.4 experimental results showed that the performance of Clarity Score indeed depends on the number of feedback documents.

In real-world situations, such a dependence on the tuning of the parameter in order to achieve meaningful performance can have adverse effects if training on one query set does not translate to another query set. Preferably, it should be possible

to set parameters automatically such that performance on the evaluation set is close to or better than the best performing parameter setting.

When computing Clarity Score, if the query language model is created from a mixture of topically relevant and off-topic documents, its score will be lower compared to a query language model that is made up only of topically relevant documents, due to the increase in vocabulary size of the language model and the added noise.

Whereas Clarity Score sets the prior to zero for documents not containing at least one query term, *Adapted Clarity* sets the prior to zero for documents not containing all m query terms, independent of the rank of the document in the result list. This effectively sets the number of feedback documents in the Clarity Score automatically; for each query, the number of feedback documents utilized in the generation of the query language model is equal to the number of documents in the collection containing all query terms.

As an example, consider TREC title topic 476: “*Jennifer Aniston*”. Among the top 1000 retrieved documents for the respective query there are 214 documents that contain both terms, 780 contain only the term *Jennifer* and 6 documents contain only the term *Aniston*. Including all documents in the query language model that do not contain both query terms adds noise to the query language model. Although documents containing only the term *Aniston* are likely to be on topic as well, the method works well as an automatic threshold. In practice, in cases where there are fewer than 10 documents fulfilling the requirement, documents with $m - 1$ query terms are included. Note that a document returned at rank i that does not contain all query terms is ignored, while a document returned at rank $j > i$ is included in the query language model if it contains all query terms.

3.5.2 Frequency-Dependent Term Selection

In Section 3.4.1 we observed that the performance of Clarity Score depends on the initial retrieval run. In the Language Modeling approach Clarity Score often performs better with retrieval algorithms relying on a small amount of smoothing. Since increased smoothing in many instances though increases the retrieval effectiveness (Table 2.2), retrieval with greater smoothing is preferred. Hence, our goal is to improve Clarity Score for retrieval runs with greater smoothing. Increased smoothing also increases the influence of high frequency terms on the KL divergence calculation (Equation 3.4), despite the fact that terms with a high document frequency do not aid in retrieval and therefore should not have a strong influence on Clarity Score. Thus, we would like to minimize the contribution of terms that have a high document frequency in the collection.

The situation is similar in a retrieval setting where we estimate a query model using feedback documents. One proposed solution by Zhai and Lafferty [169], uses expectation maximization (EM) to learn a separate weight for each of the terms in the set of feedback documents. In doing this they reduce noise from terms that are frequent in the collection, as they have less power to distinguish relevant from nonrelevant documents.

A similar approach is proposed by Hiemstra et al. [76]. The effect of both approaches is to select the terms that are frequent in the set of feedback documents, but infrequent in the collection as a whole.

Generally, a notable requirement of web retrieval is speed. Running EM to convergence, although principled, would be computationally impractical. As a remedy, to approximate the effect of selecting terms frequent in the query model, but infrequent in the collection, we select the terms from the set of feedback documents that appear in $N\%$ of the collection, where $N = \{1, 10, 100\}$. We leave the comparison of a fixed document frequency-based threshold and a variable EM-based threshold to future work.

3.6 Experiments

We tested *Adapted Clarity*, that is our adaptations on Clarity Score, on the TREC corpora and query sets already employed in Chapter 2. Apart from Clarity Score, for reasons of comparison we include a number of the best performing pre-retrieval predictor scores as already presented in the previous chapter. We also implemented four post-retrieval prediction methods, described in Section 3.2, which base their predictions on different types of information:

- *Ranking Robustness* [176] (based on document perturbation),
- *Query Feedback* [177] (based on query perturbation),
- *Normalized Query-Commitment (NQC)* [130] (based on retrieval scores), and,
- *Autocorrelation $\rho(y, \tilde{y})$ (ACSim)* [50] (based on document content).

The two parameters of the Robustness approach are the number of top ranked documents to include in the perturbation and the number of trials. We settled on 50 top documents and 100 perturbation trials; varying the parameters yielded no great changes in performance in line with the observations in [176].

The parameter settings of the Query Feedback approach were determined by training s and t on one query set and evaluating it on another. That is, the best setting of s and t on queries 301-350 was used to evaluate query sets 351-400 and 401-450; similarly for the query sets of WT10g and GOV2.

The single parameter of *NQC* is the number of top ranked documents to include in the calculation of the standard deviation. Our results are based on the top 100 documents as recommended in [130]. Similarly to the Robustness approach, small changes in the parameter do not affect the performance of this approach.

The most complex of the implemented post-retrieval predictors is *ACSim* whose parameters are the number of top ranked documents to include in the calculations, the number of most similar documents to derive the weightest sum of scores from and the similarity measure. Due to time constraints we did not train this model and instead chose the parameter settings recommended by Diaz [50] for our data sets.

In all reported experiments that follow, the smoothing parameter μ is set individually for each query set, according to the best performing retrieval effectiveness

Approach	N	TREC Vol. 4+5					WT10g			GOV2			Av.
		301-350	351-400	401-450	451-500	501-550	701-750	751-800	801-850				
AVIDF		0.591†	0.374†	0.576†	0.153	0.221	0.393†	0.315†	0.172	0.361			
SCS		0.578†	0.319†	0.518†	0.087	0.189	0.325†	0.278	0.096	0.310			
MaxSQ		0.122	0.507†	0.524†	0.429†	0.393†	0.473†	0.371†	0.306†	0.397			
MaxVAR		0.369†	0.445†	0.764†	0.381†	0.533†	0.435†	0.434†	0.345†	0.477			
AvPMI		0.316†	0.376†	0.438†	0.288†	0.235	0.431†	0.456†	0.037	0.327			
Query Feedback		0.318†	0.427†	0.382†	0.290†	0.216	0.602†	0.535†	0.490†	0.415			
ACSim		0.330†	0.536†	0.525†	0.379†	0.353†	0.550†	0.469†	0.488†	0.457			
Robustness		0.526†	0.424†	0.581†	0.312†	0.489†	0.340†	0.307†	0.415†	0.429			
NQC		0.545†	0.472†	0.678†	0.569†	0.385†	0.314†	0.297†	0.413†	0.469			
Clarity Score	100%	0.539†	0.310†	0.573†	0.260	0.251	0.603†	0.444†	0.387†	0.430			
Adapted Clarity (Fixed)	10%	<i>0.656†</i>	<i>0.409†</i>	0.572†	<i>0.348†</i>	<i>0.253</i>	0.527†	<i>0.467†</i>	<i>0.470†</i>	0.471			
	1%	0.664†	<i>0.443†</i>	<i>0.674†</i>	<i>0.545†</i>	0.199	0.527†	0.426†	0.386†	0.495			
Adapted Clarity (Automatic)	100%	<i>0.549†</i>	<i>0.485†</i>	<i>0.666†</i>	<i>0.426†</i>	<i>0.397†</i>	0.619†	0.603†	0.335†	0.519			
	10%	<i>0.629†</i>	<i>0.529†</i>	<i>0.639†</i>	<i>0.428†</i>	<i>0.366†</i>	0.577†	<i>0.602†</i>	0.356†	0.524			
	1%	<i>0.633†</i>	<i>0.511†</i>	<i>0.706†</i>	0.592†	<i>0.281</i>	0.542†	<i>0.550†</i>	0.370†	0.535			

Table 3.5: Linear correlation coefficient r with respect to the retrieval run with the best mean average precision as given in Table 2.2. Given in bold is the best performing predictor for each query set. The Adapted Clarity variations that outperform Clarity Score are given in italics. Correlations significantly different from zero are marked with † ($\alpha = 0.95$).

approach	N	TREC Vol. 4+5				WT10g				GOV2			Av.
		301-350	351-400	401-450	451-500	501-550	701-750	751-800	801-850	801-850			
AvIDF		0.314†	0.271†	0.313†	0.249†	0.187	0.277†	0.253†	0.160	0.253			
SCS		0.286†	0.227†	0.277†	0.174	0.136	0.211†	0.240†	0.095	0.206			
MaxSCQ		0.181	0.422†	0.474†	0.435 †	0.270†	0.331†	0.291†	0.209†	0.327			
MaxVAR		0.353†	0.434 †	0.494†	0.339†	0.327 †	0.288†	0.318†	0.243†	0.350			
AvPMI		0.176	0.290†	0.232†	0.208†	0.212†	0.301†	0.314†	0.034	0.221			
Query Feedback		0.294†	0.274†	0.224†	0.237†	0.160	0.432 †	0.420†	0.275†	0.290			
<i>ACSim</i>		0.332†	0.358†	0.471†	0.363†	0.265†	0.377†	0.359†	0.248†	0.347			
Robustness		0.423†	0.323†	0.424†	0.208†	0.315†	0.216†	0.199†	0.308 †	0.302			
<i>NQC</i>		0.377†	0.371†	0.381†	0.409†	0.315†	0.147	0.240†	0.255†	0.312			
Clarity Score	100%	0.420†	0.217†	0.305†	0.129	0.223†	0.415†	0.330†	0.235†	0.284			
Adapted Clarity (Fixed)	10%	<i>0.474</i> †	<i>0.304</i> †	<i>0.398</i> †	<i>0.225</i> †	<i>0.225</i> †	0.348†	<i>0.359</i> †	<i>0.291</i> †	0.328			
	1%	<i>0.485</i> †	<i>0.345</i> †	<i>0.497</i> †	<i>0.345</i> †	0.160	0.351†	0.310†	<i>0.272</i> †	0.346			
Adapted Clarity (Automatic)	100%	<i>0.423</i> †	<i>0.376</i> †	<i>0.448</i> †	<i>0.217</i> †	<i>0.286</i> †	<i>0.420</i> †	<i>0.441</i> †	0.214†	0.353			
	10%	<i>0.461</i> †	<i>0.397</i> †	<i>0.465</i> †	<i>0.260</i> †	<i>0.277</i> †	0.397†	0.457 †	0.221†	0.367			
	1%	0.500 †	<i>0.400</i> †	0.562 †	<i>0.374</i> †	0.184	0.372†	<i>0.418</i> †	<i>0.250</i> †	0.383			

Table 3.6: Kendall’s τ with respect to the retrieval run with the best mean average precision as given in Table 2.2. Given in bold is the best performing predictor for each query set. The Adapted Clarity variations that outperform Clarity Score are given in italics. Correlations significantly different from zero are marked with † ($\alpha = 0.95$).

setting (Table 2.2). Tables 3.5 and 3.6 contain the linear correlation coefficient r and Kendall’s τ respectively of the baselines, the original Clarity Score and the Adapted Clarity variations. The rows marked with *Fixed* have the same fixed number of feedback documents for all queries as well as frequency-dependent term selection. To make the results comparable with the original Clarity Score, the reported numbers are the correlation coefficients achieved with the standard setting of 500 feedback documents. The rows marked *Automatic* have their number of feedback documents set automatically as described in Section 3.5.1. The parameter N determines the amount of frequency-dependent term selection. At $N = 100\%$, all terms independent of their document frequency are included in the KL divergence calculation, at $N = 10\%$ ($N = 1\%$) only terms occurring in less than $\frac{1}{10}$ th ($\frac{1}{100}$ th) of the documents in collection are included.

The final column of Table 3.5 contains the average linear correlation coefficient over all data sets. Since r is not additive due to its skewed distribution, the average correlation does not exactly correspond to the arithmetic mean [112]. In Table 3.6 the arithmetic mean of the Kendall’s τ values are reported.

When testing the significance of the difference between the original Clarity Score and the variations of Adapted Clarity, the outcome is corpus dependent. In the case of the linear correlation coefficient r and TREC Vol. 4+5 all Adapted Clarity variations apart from one (automatic Adapted Clarity with $N = 100\%$) perform significantly better than the original Clarity Score. For the queries of the WT10g corpus only two variations significantly outperform the baseline, namely automatic Adapted Clarity with $N = 1\%$ and $N = 100\%$ respectively. No such observations can be made about the queries of GOV2: none of the Adapted Clarity variations result in a significantly higher correlation than the original Clarity Score.

The results of the significance tests for Kendall’s τ are similar. For the queries of TREC Vol. 4+5 all variations of Adapted Clarity perform significantly better than the Clarity baseline, whereas for the queries of the WT10g and GOV2 corpora none of the proposed adaptations results in a significantly better performance.

In the reported experiments in Tables 3.5 and 3.6, query 803 was removed in the evaluation of query set 801-850, as it was an extreme outlier. Due to stopword removal, the title topic “may day” is converted to the query “day”. One would expect the retrieval effectiveness of the document content based predictors of this query to be very low. The term “day” is not specific and occurs in a large number of documents. However, while the retrieval effectiveness is low (AP is 0.0) as expected, the document content based predictors assign it very high scores, in fact, Clarity Score assigns it the highest score among the 50 queries in the set by a wide margin. This surprising result can be explained when considering the makeup of the result list. We manually assessed the top 50 retrieved documents and found that the documents either contain very large HTML forms with a hundreds of different “day” options or large lists and tables, mostly filled with numbers and the term “day”. Duplicates and near duplicates lead to the outcome that 40 out of the 50 documents fall into four groups of near-duplicates, severely misleading the document content based predictors. Since the resulting correlations are then dominated by this extreme outlier we decided to remove this query from the evaluation.

3.7 Discussion

A considerable fraction of queries submitted to Web search engines occur infrequently, thus it is virtually impossible to create a representative query sample with relevance judgments to tune parameters.

For short unambiguous queries, constraining the language model to documents containing all query terms adds less noise to the language model. For terms that are ambiguous, forcing their inclusion increases noise, but this is desirable because we are capitalizing on noise in the language model to identify ambiguous queries. In the case that a query is unambiguous, but contains non-content terms, we compensate by selecting terms from the language model that are infrequent in the collection. Thus in Adapted Clarity non-content terms do not harm queries that are otherwise unambiguous.

Tables 3.5 and 3.6 show that similarly to pre-retrieval prediction methods, the performance of post-retrieval approaches also fluctuates widely over different query sets and corpora. For instance, Query Feedback achieves no significant correlation on query set 501-550 both in terms of r and τ , whereas on query set 701-750 it is among the best performing methods with $r = 0.60$ and $\tau = 0.43$ respectively. The range of predictor performance between the query sets of a single corpus is also considerable as evident for instance for Clarity Score and its correlation on query set 351-400 ($r = 0.31$) and on query set 401-450 ($r = 0.57$) respectively. The query sets of the WT10g corpus in general appear to be the most difficult for most of the evaluated post-retrieval methods to predict the query effectiveness for. Clarity Score's correlations are not significantly different from zero for both query sets, while both Query Feedback and *ACSim* perform considerably worse on WT10g's query sets than on the query sets of the other two corpora. The approach least affected by WT10g is *NQC* which exhibits moderate correlations for both query sets of WT10g. On the other hand, *NQC* performs worse than other post-retrieval approaches on GOV2, an observation we cannot explain yet and which requires further investigation.

When we consider the performance of the Adapted Clarity variations in comparison to Clarity Score we observe substantial improvements, which as pointed out in the previous section, are for some query sets large enough to be statistically significant. The largest change in correlation is observed for query set 451-500 of the WT10g corpus, where Clarity Score reaches correlations of $r = 0.26$ and $\tau = 0.13$ respectively whereas Adapted Clarity with automatically set feedback documents and $N = 1\%$ results in $r = 0.59$ and $\tau = 0.37$ respectively. When we consider the average correlation (last column in Tables 3.5 and 3.6) the Adapted Clarity variations with frequency-dependent term selection (the rows are marked as *Fixed*) outperform Clarity Score and in turn the Adapted Clarity variations with frequency-dependent term selection and automatic setting of the number of feedback documents (the rows are marked as *Automatic*) perform better than the variations with a fixed document feedback setting. With the exception of Adapted Clarity with fixed number of feedback documents and $N = 1\%$, each Adapted Clarity variation outperforms Clarity Score for at least six of the eight query sets. These observations hold for both the linear correlation coefficient and Kendall's τ . Adapted Clarity with automatically

set feedback documents and $N = 1\%$ results in the highest average correlation, indicating the benefit of both proposed adaptations.

Notable are the prediction methods that perform closest to Adapted Clarity. In the case of the linear correlation coefficient, the two prediction methods with the highest average correlation after the Adapted Clarity variations are *MaxVAR* and *NQC*, outperforming the more complex prediction methods based on document and query perturbations. With respect to Kendall's τ , *MaxVAR* is the best performing method (except for Adapted Clarity), followed by *ACSim*. This result shows, that post-retrieval approaches do not necessarily perform better than pre-retrieval prediction methods, in fact the pre-retrieval predictor *MaxVAR* outperforms all but the Adapted Clarity variations.

Predicting the quality of queries 451-550 has proven to be the most difficult across a range of predictors. In a Web environment, there are potentially millions of relevant documents for a given query. We hypothesize that the language of news articles and government websites is less varied, and the documents in these collections are more topically cohesive than Web pages. A single Web page contains a large proportion of content not related to the topic of the page itself, and furthermore even among the set of Web pages relevant to a given query, there may be a large number of different genres represented. For example in a Web setting, the set of relevant results may include pages that are largely informational (such as Wikipedia pages), pages that are largely commercial in nature, personal home pages, blogs, etcetera. Whereas the TREC Vol. 4+5 and GOV2 collections can be expected to be free of noisy pages such as spam, WT10g is not.

Furthermore, while the style for news articles is determined by a news organization and enforced to a large extent by the editors at that organization, on the Web the content is written by members of the general public with no style guidelines in place. Thus we hypothesize that one reason for the difficulty of achieving a good performance with Clarity Score on the Web corpus is the large variance in vocabulary, even among topically related documents. Since Clarity Score builds on the hypothesis that relevant documents have a more focused term distribution than non-relevant documents this metric correlates less well with noisy relevant documents (see Section 3.3.1).

3.8 Conclusions

The work reported in this chapter has focused on post-retrieval prediction algorithms. We first provided a broad overview of existing prediction methods, then focused on one particular approach: Clarity Score. Based on an analysis of Clarity Score's sensitivity to its parameter settings, we proposed two adaptations, namely setting the number of feedback documents used in the estimation of the query language model individually for each query to the number of documents that contain all query terms, and ignoring high-frequency terms in the KL divergence calculation. We evaluated these changes on three TREC test collections and compared them to a number of strong baseline approaches. We found that on average across

all evaluated query sets, *Adapted Clarity* is the best performing prediction method. Significant differences between Adapted Clarity and the original Clarity Score are only observed consistently for the queries of TREC Vol. 4+5 though. Another notable finding is the observation that the pre-retrieval predictor *MaxVAR* outperforms most evaluated post-retrieval predictors (with the exception of Adapted Clarity).