Document Quality Models for Web Ad Hoc Retrieval

Yun Zhou and W. Bruce Croft
Department of Computer Science
University of Massachusetts, Amherst
{yzhou,croft}@cs.umass.edu

ABSTRACT

The quality of document content, which is an issue that is usually ignored for the traditional ad hoc retrieval task, is a critical issue for Web search. Web pages have a huge variation in quality relative to, for example, newswire articles. To address this problem, we propose a document quality language model approach that is incorporated into the basic query likelihood retrieval model in the form of a prior probability. Our results demonstrate that, on average, the new model is significantly better than the baseline (query likelihood model) in terms of MRR and precision at the top ranks. We also give a detailed query analysis which provides some interesting insights on the limitations of the quality model and the relationship between document quality and relevance.

Categories and Subject Descriptors

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

General Terms

Experimentation

Keywords

Document quality, prior probabilities, collection-document distance, web retrieval

1. INTRODUCTION

Ad hoc retrieval is the task of finding a number of documents that are relevant to a particular information need. This task has been used as the basis for the evaluation of retrieval models since the 1960s, but it was given the name "ad hoc" first in the TREC evaluations [18]. The focus in these evaluations has been on topical relevance and, given that many of the TREC document collections consist of newswire articles, this has led to the development of retrieval models that captured topics through word distributions. For example, in the query likelihood language modeling approach [13], documents are ranked by the probability that their underlying language model can "generate" the query. Other factors relating to document content, such as the quality or genre of the text, have had very little impact. However, with the advent of the Web and test collections derived from the Web, it is clear that these other content-related document properties are much more important. In particular, due to the relative simplicity of generating and publishing web documents, the quality and style of web documents varies much more widely than the newswire-based TREC test collections. Web pages vary in quality from well-written articles to pages with very little or even no real content.

Empirical studies play an important role in IR research and many successful information retrieval (IR) systems heavily rely on the empirical tuning of model parameters. Therefore the performance of IR models typically has a close relationship with the characteristics of test collections. When the characteristics of the test collections change, as with the introduction of large Web-based collections, problems with the retrieval model can be exposed. For example, the query 'artificial intelligence' (TREC topic 741), when used to retrieve Web pages (using the query likelihood model) from the TREC GOV2 web collection [18], ranks lists of AI conferences or papers at the top, although these do not directly describe artificial intelligence at all. They are highly ranked only because the two query terms occur many times in the documents. In other words, the retrieved documents are topically relevant but are not the right type of document. In this paper, we consider this type of retrieval failure (and others described later) to be related to document "quality" and propose methods for allowing quality to influence ranking.

There has been a considerable amount of research related to Web page quality based on links. PageRank[1] and HITS[2] are two of the best-known algorithms for link structure analysis. The basic idea behind these link-based models is that a page to which many documents link is popular and therefore is likely to be of high quality. While link-based methods are clearly effective at estimating popularity, this is only one aspect of document quality. Link information has been shown to be valuable for the home-page and named-page finding TREC tasks [3,4], but participants in recent TREC web tracks [5,6,7]consistently reported that there is no conclusive benefit from the use of link information for the ad hoc task (sometimes called "content-based retrieval"). In fact, incorporating link information can sometimes even hurt retrieval performance [5,6,7].

To achieve the goal of improving the performance of Web ad hoc retrieval by exploiting document quality information, we propose a document quality model that incorporates features other than link structure. This quality model is incorporated into the basic query likelihood retrieval model in the form of a prior probability. We first show how to estimate the quality of a web document using a naïve Bayes classifier which is trained using 500 manually labeled documents from the GOV2 collection. The two features used for the classifier are information-to-noise ratio and collection-document distance. The latter is a novel feature found to be helpful for identifying low quality documents. The naïve Bayes classifier is embedded as a prior probability in the query likelihood model. We evaluate our document quality model on three TREC web collections (GOV2, WT2G and WT10G) in terms of three measures: precision at top ranked documents, mean average precision and MRR. Our results demonstrate that, on average, the retrieval model incorporating quality is significantly better than the baseline in terms of MRR and precision at the top ranks, although the impact of our model on mean average precision is quite small. Last, we give a detailed query analysis to understand the limitations of our model.

Our work offers a number of contributions. First, we propose a new document quality metric that was found to be helpful for identifying low quality documents. Second, our results show that the document quality model proposed by us can effectively improve accuracy for Web ad hoc retrieval. Third, our query analysis provides some interesting insights on the relationship between document quality and relevance.

The rest of this paper is organized as follows. Related work is discussed in section 2. In section 3, we show how the training data were created. Sections 4 and 5 describe the document quality model and the two features used for estimating the prior probability. In sections 6 and 7, we describe the experimental data and the experiments. Section 8 presents a detailed query analysis. Finally, in section 9, we summarize the main conclusions of this paper.

2. Related Work

The research on link-based approaches to the popularity aspect of quality, such as PageRank[1] and HITS[2], has already been mentioned. There have been many attempts to combine link information with content-based IR approaches to improve Web ad hoc retrieval performance [5,6,7,8,9]. However, no consistent and conclusive improvements have been demonstrated. In this paper, we focus on using content-based features rather than links to estimate document quality.

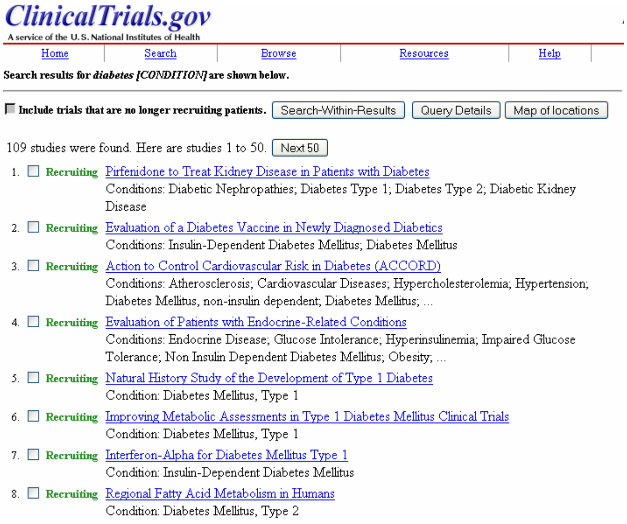
Other related work uses prior probabilities to improve language modeling based IR. The language modeling approach provides a convenient framework for incorporating prior knowledge in the form of prior probabilities. A variety of prior information, such as document length and time, has been used for ad hoc retrieval [10,11]. Kraaij et al [4] also used Web-specific features as prior knowledge for a home page retrieval system. In our approach, the prior probabilities in the language model framework are based on estimates of document quality based on content features.

There have been few attempts to directly integrate document quality into ad hoc retrieval. Zhu and Gauch [12] show that incorporating quality metrics can improve precision in a web search environment. They combine quality metrics into a vector-based algorithm in a heuristic way. The quality metrics they studied were related to currency, availability, information-to-noise ratio, authority, popularity and cohesiveness. They found information-to-noise to be possibly the most effective metric and we use this measure in our study. The other metric we use (collection-document distance) is new. One major limitation of the work is that the non-standard test collection used for evaluation is extremely small (less than 1500 documents). In fact, the test collection only comes from twenty target sites and only covers five topics. We evaluate our technique on three different Web collections that contain millions of documents.

3. Training Data

In this section, we give the details of how the training data were created. We ran 50 title queries (TREC topics 701-750) from the 2004 Terabyte Track on the GOV2 collection. The search algorithm used is the query-likelihood model with Dirichlet smoothing [13]. We looked at the top ten retrieved documents for each query (that is, 500 documents in total). We manually judged these documents either as high quality or low quality. These labeled documents will be used as the training data in our experiments described in section 7. In the experiments involving GOV2, we used five-fold cross validation to avoid testing on the training data.

Document quality is an inherently subjective concept and involves many aspects such as popularity, authority and quality of writing. Since we focus on the ad-hoc content-based retrieval task, we used the following criterion for judging a document to be low quality: A document is judged as low quality if it contains few or none of the typical sentences that would be required to describe a topic. Any other document that is not judged as low quality would be regarded as high quality. In practice, most low quality documents we found consisted of primarily tables or lists. Figure 1 gives an example of part of a typical low quality document in the training data. The document contains a list of diabetes studies that are recruiting patients for trials and was retrieved in response to the query “controlling type II diabetes”.



Search results for *diabetes [CONDITION]* are shown below.

☐ **Include trials that are no longer recruiting patients.**

109 studies were found. Here are studies 1 to 50.

1. ☐ **Recruiting** Pirfenidone to Treat Kidney Disease in Patients with Diabetes
Conditions: Diabetic Nephropathies; Diabetes Type 1; Diabetes Type 2; Diabetic Kidney Disease
2. ☐ **Recruiting** Evaluation of a Diabetes Vaccine in Newly Diagnosed Diabetics
Conditions: Insulin-Dependent Diabetes Mellitus; Diabetes Mellitus
3. ☐ **Recruiting** Action to Control Cardiovascular Risk in Diabetes (ACCORD)
Conditions: Atherosclerosis; Cardiovascular Diseases; Hypercholesterolemia; Hypertension; Diabetes Mellitus, non-insulin dependent; Diabetes Mellitus; ...
4. ☐ **Recruiting** Evaluation of Patients with Endocrine-Related Conditions
Conditions: Endocrine Disease; Glucose Intolerance; Hyperinsulinemia; Impaired Glucose Tolerance; Non Insulin Dependent Diabetes Mellitus; Obesity; ...
5. ☐ **Recruiting** Natural History Study of the Development of Type 1 Diabetes
Condition: Diabetes Mellitus, Type 1
6. ☐ **Recruiting** Improving Metabolic Assessments in Type 1 Diabetes Mellitus Clinical Trials
Condition: Diabetes Mellitus, Type 1
7. ☐ **Recruiting** Interferon-Alpha for Diabetes Mellitus Type 1
Condition: Insulin-Dependent Diabetes Mellitus
8. ☐ **Recruiting** Regional Fatty Acid Metabolism in Humans
Condition: Diabetes Mellitus, Type 2

Figure 1: “Low quality” document retrieved in response to the query *controlling type II diabetes*.

The intuition behind this basis for quality judgments is that a relevant document for the TREC ad hoc task usually explains or describes some topic using sentences with typical English structure and vocabulary. Therefore, documents like tables or lists are unlikely to be relevant for ad hoc queries.

We examined the relationship between relevance and document quality and Table 3.1 shows the distribution of relevant documents over two classes: high quality and low quality documents.

Table 3.1 Distribution of relevant documents in the training data

	Relevant	Non-relevant
High quality	238	171
Low quality	9	82

As we can see, the proportion of relevant documents among low quality documents is much lower than that in high quality ones. In section 8, we examine cases where low quality documents may be relevant. Overall, based on our training data, successfully recognizing low quality documents should be helpful for improving retrieval performance.

4. Quality Metrics

Our approach depends on the identification of metrics or document features that are predictive of quality. As mentioned previously, we focus on two metrics, collection-document distance and information-to-noise ratio, the first of which is new and the second having been used with some success in a previous study [12]. Although the results are not reported here, we have tried a number of other content-based features, such as document length, and the mean and variance of the document word distribution, but found these to be not as predictive of quality.

4.1 Collection-document distance

The Collection-Document Distance (CDD for short), is simply the relative entropy, or Kullback-Leibler (KL) divergence, between the collection and document unigram language models. The collection or background language model is estimated using the word occurrence frequencies over the whole collection (e.g. GOV2). A similar measure, called Clarity [19], has been used to predict which queries will perform well.

Given a document D and a collection C, the CDD is given by

$$CDD = \sum_{w} P_{coll}(w \mid C) \log \frac{P_{coll}(w \mid C)}{P_{doc}(w \mid D)} \qquad (4.1)$$

$$where \quad P(w \mid D) = \lambda P_{doc}(w \mid D) + (1-\lambda) P_{coll}(w \mid C)$$

$$P_{doc}(w \mid D) = \frac{\#Count(w,D)}{\| D \|}, P_{coll}(w \mid C) = \frac{\#Count(w,C)}{\| C \|}$$

In this formulation, we use linear smoothing for estimating the document language model probabilities.

Our hypothesis is that low quality documents will have unusual word distributions. In other words, if a document differs significantly from the word usage in an average document, the quality of this document may be low. In the CDD measure, the average document is represented by the collection language model. The KL divergence between the collection language model and the document language model (i.e. the CDD) indicates how different these distributions are. The higher the CDD is, the more unusual the word distribution of the document is, and the more likely, according to our hypothesis, that the document is of low quality.

Let us consider three cases that are helpful for understanding why CDD can predict low quality documents.

Case 1: documents that are tables or lists. Common words, such as pronouns, adjectives and verbs, would have very low numbers of occurrences, which makes the document language models quite different from the collection language model.

Case 2: documents that have misspelled words. The probability of a misspelled word in the collection is much lower than that of normal words. If any document contains misspelled words, the CDD tends to be high.

Case 3: documents where the frequency of some term is unnecessarily high. Since the web environment contains competing profit seeking ventures, one may intentionally increase the occurrence of some keywords in a document to get attention. CDD can recognize this case.

As an alternative to Equation 4.1, one can compute the divergence with the role of the collection language model and the document language model reversed. The method shown in Equation 4.1 performs slightly better in our evaluations and is used throughout this paper. The value of the parameter λ is determined empirically and is 0.8 for all runs in this paper.

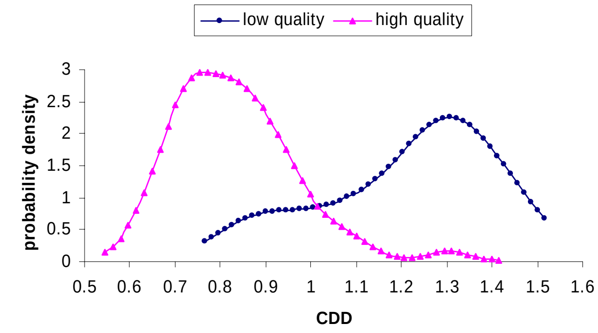


Figure 2: Distribution of CDD values for low and high quality documents.

Fig 2 shows the distributions of CDD values for high quality and low quality documents respectively. These two distributions are estimated from our training data by the Kernel density estimation method that will be discussed in the next section. We can see that there is an obvious separation between the two classes of documents.

4.2 Information-to-noise ratio

The information-to-noise ratio is computed as the total number of terms in the documents after indexing divided by the raw size of the document [12]. This metric predicts low quality documents based on a different characteristic than the CDD metric. Consider a web document that has only a few words and many HTML tags which will be removed after indexing. The information-to-noise ratio of this document is very low and the quality of this document also tends to be low.

Fig 3 shows the distributions of information-to-noise ratios for high quality and low quality documents respectively. The two distributions are also estimated from our training data by the Kernel density estimation method. As we can see, a document with a low information-to-noise ratio is much more likely to be of low quality.

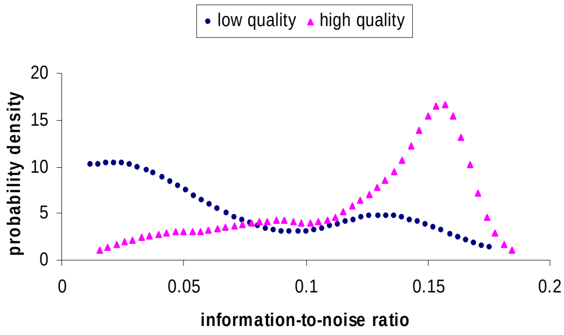


Figure 3: Distribution of information-to-noise ratios for low and high quality documents.

5. Document Quality Language Model

Our retrieval model, which we call the document quality language model, is built on the top of the basic query likelihood model by incorporating document quality metrics in the form of a prior probability. Specifically, we estimate the quality of a Web document by a naïve Bayes classifier that is embedded as a prior probability in the query likelihood model.

5.1 Review of query-likelihood language model

The query likelihood model [13] is a simple and robust language model approach to retrieval and is used as the baseline in this paper. In this model, given a query Q and a document D, the document D is ranked by P(D|Q), the probability that D is relevant given Q. By Bayes rule, we have

$$P(D \mid Q) \propto P(D)P(Q \mid D) \qquad (5.1)$$

P(Q|D) is the probability that the document D can "generate" the query Q . If we assume that each document is assigned a multinomial distribution over words and use linear smoothing[1], P(Q|D) can be calculated as :

$$P(Q \mid D) = \prod_{w \in Q} P(w \mid D) = \prod_{w \in Q} \lambda P_{doc}(w \mid D) + (1-\lambda) P_{coll}(w \mid C) \qquad (5.2)$$

$$where \quad P_{doc}(w \mid D) = \frac{\#Count(w,D)}{\| D \|}, P_{coll}(w \mid C) = \frac{\#Count(w,C)}{\| C \|}$$

$$C \quad denotes \quad collection$$

P(D) is the document prior probability and usually assumed to be uniform in the query-likelihood model. By doing so, documents are actually ranked solely by P(Q|D). Even though relevance is not explicitly mentioned in formula 5.1, P(D) can be viewed as prior knowledge about the relevance of document D [14]. As mentioned previously, various factors such as document length and currency have been used to derive estimates for this prior probability of relevance. In this paper, we focus on quality-related estimates for this prior, although these could be combined with other estimates for a more general prior.

5.2 Document quality language model

In section 3 we divided all documents into two types: high quality and low quality .We also showed that high quality documents are more likely to be relevant than low quality ones. Therefore, we propose to replace the prior probability in formula 5.1 by the probability that the quality of the document is high given the two metrics described in the last section.

Let D denote a document. Note that we assume that all documents belong to one of the two classes: high quality and low quality. Let H denote the high quality class, L denote the low quality class, X denote a vector of quality metric values, and π_H and π_L denote the prior probabilities of the high quality class and the low quality class respectively. Let f_H and f_L denote the probability density functions of the high quality class and the low quality class respectively. By Bayes rule , we have:

$$\Pr(D = H \mid X = x_0) = \frac{\pi_H f_H(x_0)}{\pi_H f_H(x_0) + \pi_L f_L(x_0)} \qquad (5.3)$$

Given multiple features (in this case, quality metric values) it is common to assume independence among the features. In fact, we examined the training data and found there is little correlation between the two metrics. Under this assumption, we have

$$f_j(X) = f_j(x_0) f_j(x_1), \quad j = H, L \quad (5.4)$$

where x_0 is the CCD metric and x_1 is the information-noise ratio.

5.3 Kernel density estimation

The key part of computing Pr(D=H|X) is the estimation of the probability density functions in Equation 5.3, since π_H and π_L can be simply estimated by the relative frequencies in the training data. However, it is not easy to estimate these functions since we do not know what distribution the two metrics actually follow. Instead, we adopt Kernel density estimation which does not assume any specific distribution on the features we want to estimate. Kernel density estimators belong to a class of estimators called *non-parametric* density estimators that have no fixed structure and depend upon all data points to reach an estimate.

Assume we have a random sample x_1, x_2, …x_N drawn from a probability density function *f(x)* and we wish to estimate *f(x)* at a point x_0 , the Kernel density estimator for *f(x)* at the point x_0 is defined as [14]:

$$\hat{f}(x) = \frac{1}{N\lambda} \sum_{i=1}^{N} K_{\lambda}(x_0, x_i) \quad (5.5)$$

where λ is the bandwidth and K_λ is a Kernel function.

In this paper we use the Gaussian Kernel and Equation 5.4 can be written as

$$\hat{f}(x) = \frac{1}{N\lambda\sqrt{2\pi}} \sum_{i=1}^{N} \exp(-\frac{(x_i - x_0)^2}{2\lambda^2}) \quad (5.6)$$

There is a standard way to select the bandwidth (λ) based on minimizing the expected square error between the estimated density and the original density [15]. In this paper, we adopt this method to calculate λ.

5.4 Summary

To summarize this section, the document quality language model is as follows:

[1] Although in our experiments we use Dirichlet smoothing.

$$P(D \mid Q) \propto P(Q \mid D)P(D = H \mid X)$$

Where P(Q|D) is the query likelihood model computed in Equation 5.2 and P(D=H|X) can be computed by Equations 5.3, 5.4 and 5.6. π_H and π_L are estimated by the relative frequencies in the training data.

6. Experimental data

Our document quality model was evaluated on a variety of web test collections. (see table 6.1)

Table 6.1: Summary of test collections

Test collection	Size (GB)	number of docs	TREC topics
GOV2	426	25,205,179	701-750
WT10G	11	1,692,096	501-550
WT2G	2	247,491	401-450

WT10G is a subset of the VLC collection, which is a subset of a 1997 crawl of the Web. WT2G is a subset of WT10G. GOV2 consists of a crawl of the .gov web domain. Most of the documents in the GOV2 collection are HTML documents but some of them are plain text, PDF, PS, and MS Word documents. More information about the three test collections can be found at [16].

Given the fact that short keyword queries dominate current web search engines, the queries used in our experiments are from the title field TREC topic 401-450,501-550 and 701-750 as shown in the fourth column of table 6.1.

As described in section 3, we manually labeled 500 documents and use them as the training data for our model. For the runs on the GOV2 collection we split all training data into five parts according to which query the document is for. Then we did a five-fold cross validation with one part reserved for testing and the rest used for training. For the runs on the WT2G and WT10G collections, we used all the training data they are different collections.

7. Results

In this section we present the results of comparisons between the document quality model and the query likelihood model on the three Web test collections. Three metrics are used for evaluation: precision at top retrieved documents, mean average precision and MRR (mean reciprocal rank) . Our results show that the document quality model significantly outperforms the baseline in the evaluation using precision at top ranked documents and MRR, although the differences on MAP between the two models are quite small.

For query likelihood retrieval, we use Dirichlet smoothing with a smoothing parameter of 2500 for all runs.

7.1 Results for precision at top ranks

In a typical Web search environment, few people would look at more than the first ten or twenty results. Precision at the top ranks is a very important metric since it reflects the concern with high retrieval accuracy. In this paper, we evaluate precision at 4 rank levels: 5, 10, 15 and 20.

Table 7.1 shows the precisions at top ranks on the GOV2 collection. To better compare our model with the baseline, all queries are divided into three types: “Pos”, “Neg” and “Eq”, which means our model is better, worse or equal to the baseline respectively. The last column in table 7.1 shows the numbers of the three types of queries.

Table 7.1: Precision on the GOV2 collection. “Pos” means result is better than the baseline, “Neg” means result is worse than the baseline, “Eq” means result is the same as the baseline.

Precision @	Query-likelihood model	Document quality model	Pos.	Neg.	Eq.
5 docs	0.5184	0.5633	11	6	32
10 docs	0.4980	0.5306	12	7	30
15 docs	0.4653	0.5088	18	6	25
20 docs	0.4612	0.5020	19	7	23

We can see that the document quality model consistently outperforms the baseline at all of the 4 rank levels. On the other hand, the majority of the queries are not affected by the quality-based prior. One reason is that high quality documents, where the differences in the prior probabilities tend to be negligible, consist of a large part of the whole collection. In other words, our model can make a difference only when there are enough low quality documents in a rank list. Approximately twice as many queries are improved by this technique than are hurt.

The results for WT2G and WT10G are shown in table 7.2 and 7.3 respectively. As with the results on the GOV2 collection, the document quality model consistently improves precision. Moreover, considering the limited size of the training data, we believe that performance could be further improved by including more training data from a variety of web collections.

Table 7.2: Precision on the WT2G collection

Precision @	Query-likelihood model	Document quality model	Pos.	Neg.	Eq.
5 docs	0.4960	0.5240	9	3	38
10 docs	0.4640	0.4760	10	4	36
15 docs	0.4107	0.4280	10	3	37
20 docs	0.3880	0.3920	10	7	33

When combining all queries together from the three test collections, Table 7.4 shows the numbers of queries that are better or worse than the baseline respectively. (We ignore the case where the two have the same performance). Here we did a Fisher sign test [17] with 95% confidence interval to verify that the differences are significant.

In summary, these results suggest that incorporating document quality information can significantly improve precision at the top ranks.

Table 7.3: Precision on the WT10G collection

Precision @	Query-likelihood model	Document quality model	Pos.	Neg.	Eq.
5 docs	0.3440	0.3640	9	6	35
10 docs	0.3000	0.3240	13	5	32
15 docs	0.2880	0.2907	13	12	25
20 docs	0.2660	0.2900	19	9	22

Table 7.4: Comparison of the two models on all queries

Precision @	Pos	Neg	Statistically significant ?
5 docs	29	15	Yes
10 docs	35	16	Yes
15 docs	41	21	Yes
20 docs	48	23	Yes

7.2 Mean average precision results

Mean average precision (MAP for short) is the most frequently used measure for ad hoc retrieval. In our view, MAP is a less important measure than precision at the top ranks for a typical web user, but to fully evaluate and understand the quality model, we include this measure.

Table 7.5 shows the mean average precision on GOV2, WT2G and WT10G. Percentage improvements with respect to the baseline are also given. “Pos”, “Neg” and “Eq” have the same meaning mentioned in section 7.1. As we can see, the differences between the two models are small, which suggests that on average there is no significant positive or negative benefit from the quality-based prior in terms of MAP.

In section 8, we will discuss the reasons for this in more detail.

Table 7.5: MAP on the three test collections

Collection	Query-likelihood model	Document quality model	Pos.	Neg.	Eq.
GOV2	0.2516	0.2493 (-0.9%)	20	30	0
WT2G	0.3135	0.3220 (+2.7%)	34	16	0
WT10G	0.1831	0.1840 (+ 0.5%)	27	21	2

7.3 MRR results

MRR, which is defined as the inverse of the rank of the first relevant document, is another way for evaluating high accuracy retrieval. MRR is useful in cases where users are primarily looking for one correct answer and want that answer ranked as high as possible. Home-page finding and question answering are two TREC tasks where MRR is the standard for evaluation.

Table 7.6 shows MRR on the three test collections. Since the document quality model is consistently better than the baseline in terms of precision at top ranked documents, it is not surprising that our model performs better on GOV2 and WT2G. WT10G is an exception where our model is slightly worse than the baseline. We discuss this more in the next section, but MRR is more sensitive to a small fluctuation in the rank list than precision. For example, if the rank of the first relevant document is second instead of first, the MRR drops from 1.0 to 0.5 while the precision in the top 5 documents may change very little.

The results given in Table 7.7 are the average of the MRR when putting all queries together. Again we did a Fisher sign test with 95% confidence interval. In summary, these results indicate that the document quality model is significantly better than the baseline in terms of MRR.

Table 7.6: MRR on the three test collections

Collection	Query-likelihood model	Document quality model	Pos.	Neg.	Eq.
GOV2	0.7096	0.7746	15	5	29
WT2G	0.7406	0.7781	9	1	40
WT10G	0.6215	0.6139	9	8	33

Table 7.7: MRR comparison of the two models on all queries

Baseline	Document quality model	Pos.	Neg.	Statistically significant?
0.6906	0.7222	33	14	Yes

8. Query Analysis

In section 7 we showed that on average the quality-based model can effectively improve precision and MRR with respect to the baseline. To understand the limitations of this approach and potentially find improvements, we analyzed the queries for which the model did most poorly in terms of MAP and MRR. Below are the details of six of these queries and our explanations why the quality model does not perform well. Note that in the cases where no MRR is listed, the two models have the same MRR value.

Example 1 (GOV2):

Query	Topic	Baseline	Quality model
Nuclear reactor types	748	0.1138 (MAP) 1.0 (MRR)	0.0628 (MAP) 0.5 (MRR)

Explanation: According to the narrative for this topic, relevant documents only need to mention the names of the types of nuclear reactor power plants. Therefore, low quality documents like lists or tables could be relevant for this topic. The quality model penalizes some of these relevant documents.

Example 2 (GOV2):

Query	Topic	Baseline	Quality model
Green party political views	704	0.1733 (MAP)	0.0662 (MAP)

Explanation: it seems that low quality documents are not likely to be relevant for this topic. However, the narrative section of this topic says “Any members’ names noted are considered relevant”. There are a few low quality documents judged as relevant only because the names of green party members are listed, which leads to the failure of our model in this case.

Example 3 (WT10G):

Query	Topic	Baseline	Quality model
History of skateboarding	506	0.1276 (MAP) 0.25 (MRR)	0.017 (MAP) 0.026 (MRR)

Explanation: There are only two documents judged as relevant for this topic. One of the two is retrieved by neither the quality model nor the baseline. The other one that is highly ranked by the baseline is a list.

Example 4 (WT10G):

Query	Topic	Baseline	Quality model
Instruments to forecast the weather	541	0.2588 (MAP)	0.1497 (MAP)

Explanation: As in example 1, low quality documents such as lists can be relevant documents for this topic

Example 5 (WT2G):

Query	Topic	Baseline	Quality model
Cuba sugar exports	414	0.5898 (MAP) 1.0 (MRR)	0.4806 (MAP) 0.5 (MRR)

Explanation: In the description section of this topic it says "How much sugar does Cuba export and which countries import it". As we can see, just numbers and names are enough to be relevant for this topic.

Example 6 (WT2G):

Query	Topic	Baseline	Quality model
Quilts, income	418	0.3643 (MAP)	0.2634 (MAP)

Explanation: The narrative section of this topic states "Documents mentioning quilting books, quilting classes, quilted objects and museum exhibits of quilts are all relevant". According to these criteria, low quality documents can be relevant for this topic.

In summary, it seems that the biggest problem for the current quality model is that there are queries with relevant documents that are low quality according to the model. To better understand this issue, we manually divided all queries for the GOV2 collection into the following two types:

Type one: queries that are not likely to have relevant low quality documents.

Type two: queries that are likely to have relevant low quality documents.

According to our classification, there are 33 type one queries and 16 type two queries. The heuristic we used for the classification is that if a few named entities are enough to satisfy the information need as defined in the narrative, the query will be classified as type two. Otherwise, if detailed topic description is needed, the query will be classified as type one. Of course, the classification is still somewhat ambiguous for some queries.

Table 8.1 shows MAP results for the two types of queries defined above. Percentage improvements with respect to the baseline are also given. Considering the explanations given above, it is not surprising to see that the performance of the document quality model is quite low on the type two queries. On the other hand, our model is better than the baseline on the type one queries, although the improvement is small. This is because there are relatively few low quality documents in a typical ranked list and MAP is based on the whole ranked list. Another interesting observation from table 8.1 is that both two models performance better on type one queries. Even though we currently can not automatically distinguish the two types of queries, our analysis suggests that a different strategy is needed to improve the performance of type two queries.

Table 8.1: MAP on the two types of queries

Query Type	Query likelihood model	Document quality model
Type One	0.2664	0.2710 (+1.7%)
Type Two	0.2209	0.2045 (-7.4%)

9. Conclusions and Future Work

We presented the document quality language model for improving the performance of Web ad hoc retrieval. The model provides a framework for integrating quality metrics into the language modeling approach.

We showed on a variety of TREC Web test collections that the new model can improve precision at the top ranks and MRR by penalizing low quality documents, although there is no significant improvement on MAP. Additionally, we observed that relevant documents could be of low quality for some queries.

In the future, we will explore more features related to quality. We also plan to further investigate the relationship between quality and relevance, and potentially develop different priors for different types of queries.

10. Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by SPAWARSYSCEN-SD grant number N66001-02-1-8903 and in part by NSF grant #IIS-0527159.

Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

11. References

[1] S.Brin and L.Page, The anatomy of a large-scale hypertextual web search engine, *Computer Network and ISDN Systems*, 30(1-7):107-117,1998.

[2] J.M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604-632,1999.

[3] T. Westerveld, W.Kraaij, and D.Hiemstra, Retrieving web pages using contents, links, URL's and anchors. *the Tenth Text Retrieval Conferences* (*TREC-2001)*, 52-61, NIST, 2002

[4] K.Wessel, W.Thijs and H.Djoerd, The importance of prior probabilities for entry page search, *Proceedings of SIGIR 2002,* 27-34, 2002.

[5] D.Hawking, Overview of the TREC-9 web track. *The Ninth Text Retrieval Conferences* (*TREC9)*, 87-102, NIST,2001

[6]D.Hawking ,E.voorhees,N.Craswell,and P.Bailey, Overview of the TREC-8 web track. *The Eighht Text Retrieval Conference* (*TREC8),*131-148, NIST, 2000

[7] D.Hawking and N.Craswell, Overview of the TREC-2001 web track. *The Tenth Text Retrieval Conferences* (*TREC-2001),* 25-31 NIST, 2002

[8] K.Yang, Combing text and link-based retrieval method for web IR. In *the Tenth Text Retrieval Conferences* (*TREC-2001)*, 609-618, NIST, 2002.

[9] W. Kraaij and T. Westerveld, TNO-UT at TREC-9: How Different are Web Documents? In *the Ninth Text Retrieval Conferences* (*TREC9)*, 665-672, NIST, 2001.

[10] D.Hiemstra and W.Kraaij. Twenty-one at TREC-7: Ad-hoc and cross-language track. In *the Seventh Text Retrieval Conferences* (*TREC7),* NIST, 1999.

[11] X. LI and W.B. Croft, Time-based language models. In *proceedings of the Twelfth International Conference on Information and Knowledge Management* (CIKM'03), 2003.

[12] X. Zhu and S. Gauch, Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. In *Proceedings of SIGIR 2000* , 288-295, 2000.

[13] W.B. Croft and J. Lafferty (eds.), *Language Models for Information Retrieval.* Kluwer Academic Publishers, 2003.

[14] T. Hastie, R. Tibshirani, J. H. Friedman .The Elements of Statistical Learning, Section 6, Kernel Method. Springer press, 2001

[15] Kernel density estimation, http://www.xplore-stat.de/tutorials/smoothernode2.html

[16] Web Research Collections, http://es.csiro.au/TRECWeb/

[17] Eric W. Weisstein, Fisher Sign Test. From MathWorld--A Wolfram Web Resource,

http://mathworld.wolfram.com/FisherSignTest.html

[18] Text REtrieval Conference (TREC) Home Page, http://trec.nist.gov/

[19] S. CroneTowsend, Y. Zhou, and W.B. Croft, "Predicting query performance", *Proceedings of SIGIR 2002*, 299-306, 2002.