Document Quality Models for Web Ad Hoc Retrieval[1]

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 precision at the top ranks.

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

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 is incorporated into the basic query likelihood retrieval model in the form of a prior probability. Instead of relying on hyperlink analysis, we use two content features to estimate Web document quality. One of the two features is a novel document quality metric that was found to be helpful for identifying low quality documents.

2. Document Quality Language Model

The first step of our approach depends on the identification of metrics or document features that are predictive of quality. In this paper 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 [1] where information-to-noise ratio is simply defined as the total number of terms in the documents after indexing divided by the raw size of the document. We now show how to compute the first metric, collection-document distance.

Given a document D and a collection C, the Collection-Document Distance (CDD for short) is given by

$$CDD = \sum_{w} P_{coll}(w \mid C) \log \frac{P_{coll}(w \mid C)}{P(w \mid D)} \qquad (1)$$

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

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

The basic idea behind CDD comes from the observation that documents like tables or lists are unlikely to be relevant for ad hoc queries because a relevant document for the TREC ad hoc task usually explains or describes some topic using sentences with typical English structure and vocabulary. Therefore, we hypothesize that such 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. Figure 1 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[3].

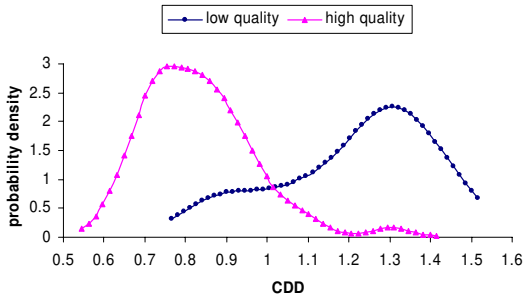

Figure 1: Distribution of CDD values for low and high quality documents

Next we show how to estimate the quality of a Web document by a naïve Bayes classifier combining the two quality metrics mentioned above. Let D denote a document, Let H denote the high quality class, L denote the low quality class, X denote a

ACM 1-59593-140-6/05/0010 . A full version of this paper is in [3].

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:

$$\Pr(D = H \mid X = x_0) = \frac{\pi_H f_H(x_0)}{\pi_H f_H(x_0) + \pi_L f_L(x_0)} \quad (2)$$

By assuming independence among the features, we have

$$f_j(X) = f_j(x_0) f_j(x_1), \ j = H, L \quad (3)$$

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

π_H and π_L in Equation 2 can be simply estimated by the relative frequencies in the training data(please refer to [3] for the details of our training data). To estimate f_H and f_L in Equation 3 from the training data, we adopt the Kernel density estimation and choose the Gaussian Kernel. Without loss of generality, assume we have a random sample $x_1, x_2, \ldots x_N$ drawn from a probability density function *f(x)* and we wish to estimate *f(x)* at a point x_0 , the Guassian Kernel density estimator for *f(x)* at the point x_0 is defined as [5]

$$\hat{f}(x) = \frac{1}{N\lambda\sqrt{2\pi}} \sum_{i=1}^{N} \exp\left(-\frac{(x_i - x_0)^2}{2\lambda^2}\right) \quad (4)$$

Finally, the probability given in Equation 2 is embedded as a prior probability in the query likelihood model. Specifically, given a query Q and a document D, let P(D|Q) be the probability that D is relevant given Q, the document quality language model is as follows:

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

Where P(*Q*|*D*) is the query likelihood model described in [2] and P(*D*=*H*|*X*) computed by Equation 2 can be interpreted as the document prior probability that reflects prior knowledge about the relevance of the document D[4].

3. Results

Table 1 shows the precision at top ranks on the GOV2 collection. The queries used are from the title field TREC topic 701-750.The retrieval parameter settings are given in [3]. We did

Table 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. Bold cases means the results are statistically significant

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

Table 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

Table 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

a Fisher sign test with 95% confidence and the bold numbers mean the results are statistically significant.To better compare our model with the baseline query-likelihood model, 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 1 shows the numbers of the three types of queries. The results for WT2G and WT10G are shown in table 2 and 3 respectively. For these two collections, we used the title field TREC topic 401-450 and 501-550 as queries.

4. 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.

5. References

[1] 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.

[2]F.Song and W.B. Croft. A general language model for information retrieval. In *Proceedings of SIGIR 1999* , 279-280, 1999

[3]Y.Zhou and W.B.Croft Document quality models for web adhoc retrieval Technical report IR-432, Center for Intelligent Information Retrieval, University of Massachusetts,2004

[4] W. Kraaij and T. Westerveld and D. Hiemstra, The importance of prior probabilities for entry page search, *Proceedings of SIGIR 2002,* 27-34, 2002.

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