TNO/UT *at TREC-9: How different are Web documents?

[†]Wessel Kraaij, and [††]Thijs Westerveld

[†]TNO-TPD
P.O. Box 155, 2600 AD Delft
The Netherlands
kraaij@tpd.tno.nl

[††]University of Twente, CTIT
P.O. Box 217, 7500 AE Enschede
The Netherlands
westerve@cs.utwente.nl

Abstract

Although at first sight, the web track might seem a copy of the ad hoc track, we discovered that some small adjustments had to be made to our systems to run the web evaluation. As we expected, the basic language model based IR model worked effectively on this data. Blind feedback methods however, seem less effective on web data. We also experimented with rescoring the documents based on several algorithms that exploit link information. These methods yielded no positive result.

1 Introduction

The basic idea for the web track run was to modify our ad-hoc system for the main web task and perform some experiments with the link structure information. We did not know to what scale we would have to re-engineer our systems to be able to deal with 10 giga bytes of data which is about 5 times larger than the ad-hoc collection. We applied the same IR model based on an interpolated unigram language model which had proven to be succesful on several data collections and tasks: Ad hoc, CLIR, SDR and filtering. The model will be presented in Section 2.

2 Retrieval Model

All runs were carried out with an information retrieval system based on a simple unigram language model. The basic idea is that documents can be represented by simple statistical language models. Now, if a query is more probable given a language model based on document d_1, than given e.g. a language model based on document d_2, then we hypothesise that the document d_1 is more relevant to the query than document d_2. Thus the probability of generating a certain query given a document-based language model can serve as a score to rank documents with respect to relevance.

$$P(T_1, T_2, \cdots, T_n|D_k)P(D_k) \;=\; P(D_k)\prod_{i=1}^{n}\lambda P(T_i|D_k) + (1-\lambda)P(T_i) \tag{1}$$

In the above formula, $T_1, .., T_n$ represents a query, $P(D_k)$ is the a priori probability of relevance of a document. The product term consists of the probability of a term given a document $P(T_i|D_k)$ interpolated with the marginal $P(T_i)$. For the a priori $P(D_k)$ we usually take:

$$P(D_k) = \frac{\text{dlen}_k}{\sum_{j=1}^{N}\text{dlen}_j} \tag{2}$$

*The TNO and University of Twente team is a continuation of the "TwentyOne" cooperative team which participated in previous TREC evaluations.

This choice can be motivated by the fact that empirical studies by Singhal [4] have shown that there is usually a linear relationship between probability of relevance and document length.

The model was implemented in a vector product form supported by the TNO search engine. Our system was able to index the 10 Gigabyte dataset in roughly 20 hours on a SUN ultrasparc 300 Mhz. No re-engineering was necessary, except for the HTML entity conversion, which broke on several non-conforming documents.

3 Content only Experiments

We experimented with several variants for the estimator of the marginal $P(T_i)$ in formula (1). We compared an estimator based on the document frequency:

$$P(T_i) = \text{df}_i / N \tag{3}$$

with an estimator based on the collection frequency:

$$P(T_i) = \sum_{j=1}^{N} \text{tf}_{ij} / \sum_{i=1}^{V} \sum_{j=1}^{N} \text{tf}_{ij} \tag{4}$$

and an estimator based on the term frequency averaged over all documents:

$$P(T_i) = \sum_{j=1}^{N} (\text{tf}_{ij} / \text{dlen}_j) \tag{5}$$

In these estimators, N is the number of documents and V the indexing vocabulary.

A second experiment dealt with score normalisation. Score normalisation is not necessary for the web task, but is relevant for other tasks like CLIR and topic tracking. We had found that dividing the RSV by the query length helps to normalize scores across topics. This makes sense because the RSV is composed of a sum of log terms. (cf. [3] for a description of the vector space implementation of the model, which is based on taking the log of the probability, thereby converting the product into a summation) However, when we choose a model which includes a document prior $P(D_k)$, the RSV is not a sum of query term related addends anymore, because the document prior is a constant probability, independent of the query length, which is even added when the query has zero length. We assumed that we could correct for this problem by assuming that both the document prior and the query dependent score component (the first term) are independent sources of evidence, in that case we can add a weighting component β , which controls the ratio of the prior evidence component in the final RSV.

$$RSV(Q, D_k) = 1/T_n \sum_{T_i = T_1} T_n \log(\lambda_i P(T_i|D_k) + (1 - \lambda_i) P(T_i)) + \beta \log P(D_k) \tag{6}$$

Experiments showed however, that the assumption that both sources of evidence are independent, is not true. The original model where the document prior is seen as an internal component of the model and where the sum component is not normalised separately showed the best performance. This leaves the RSV normalisation problem (which is not relevant for the web task) yet unsolved. We hypothesize that the document priors are especially helpful as an additional probabilistic knowledge source, when the system does not have a lot of information about the topic of interest (e.g. the query is short). For more informative queries, the influence of the a priori knowledge that longer documents tend to be more often significant is small, because this effect is implicitly coerced by the retrieval model. The longer the query, the lower the probability that a short document contains all query terms.

We tested several blind feedback methods on the TREC8 2 Gigabyte small web task. We did not find a consistent improvement, for title queries the performance was even hurt. We decided to refrain from feedback in the TREC9 web runs. We think the blind feedback was especially troubled by the presence of typos, which are abundant in web documents. These typos receive a high weight in most pseudo feedback strategies, because of there low document frequency. A more detailed analysis is required to study whether this is the only problem.

Table 1 gives the results of the content only runs. We have focussed on title only runs, because we feel these are most real-life and challenging.

runtag	official run	description	average precision
tnout9t2	yes	title run with 0.5 doc priors	0.1801
tnoutf1	yes	full run without doc priors	0.2178
df-estimator	no	title with doc priors	0.1871
df-estimator	no	title without doc priors	0.1465
cf-estimator	no	title with doc priors	0.1884
avtf-estimator	no	title with doc priors	0.1871
df-estimator	no	full with doc priors	0.2240

Table 1: Content-only results

The first of the official runs (tnout9t2) is a title run based on the third (average tf) estimator with a lambda value of 0.01 to enhance coordination, a prior weight β of 0.5 while dividing the first term by the query length (according to formula (6)), the second official run (tnoutgf1) is a full run based on the first estimator using the standard model of (1) without document priors. In the full run terms from the title receive a triple weight, terms from the description run receive a double weight and terms from the narrative section a single weight. This choice was motivated by some post hoc experiments on prior collections.

We have done some additional experiments. First we modified our tokenizer to allow query terms with digits to enter the fuzzy matching process. This brought a small but insignificant improvement (only one topic changed).

We also re-tested the different estimators in combination with standard document priors and different lambda values. It turned out that the choice of a lambda value of 0.80-0.90 was best for all three estimators, with very small performance differences the table shows results for lambda=0.1. The second estimator, based on the collection frequencies scored best, but practically spoken, the three estimators work about as well.

We have made some additional plots to check whether the assumption that probability of relevance is linearly correlated with document length holds for a number of collections:

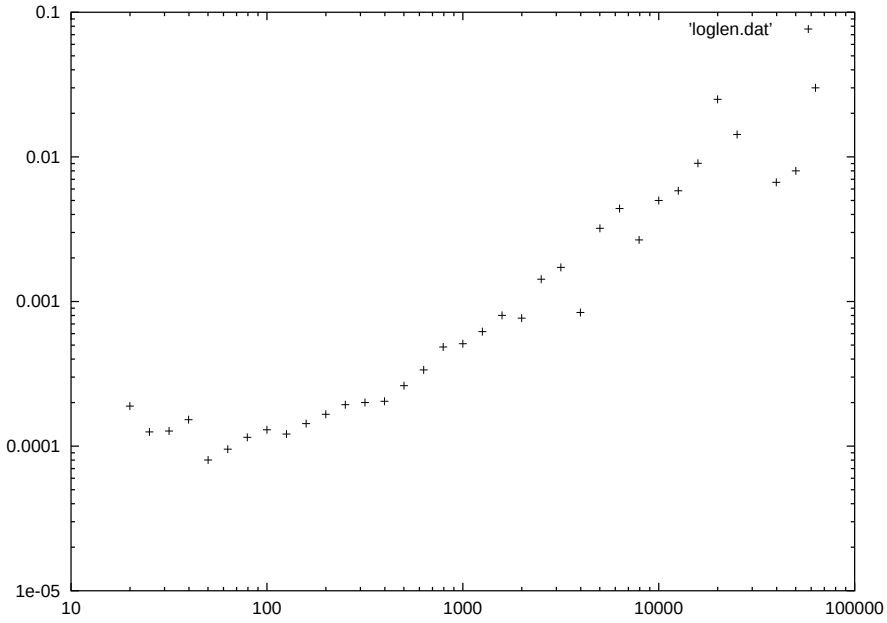

Figure 1: $P(D|l)$ for the TREC7 Ad Hoc collection

Figures 1,2, 3 and 4 show plots of $P(D_k \text{isRel} | \text{dlen}_k \in \text{bin}_k)$. Similar to Singhal, we binned the documents from

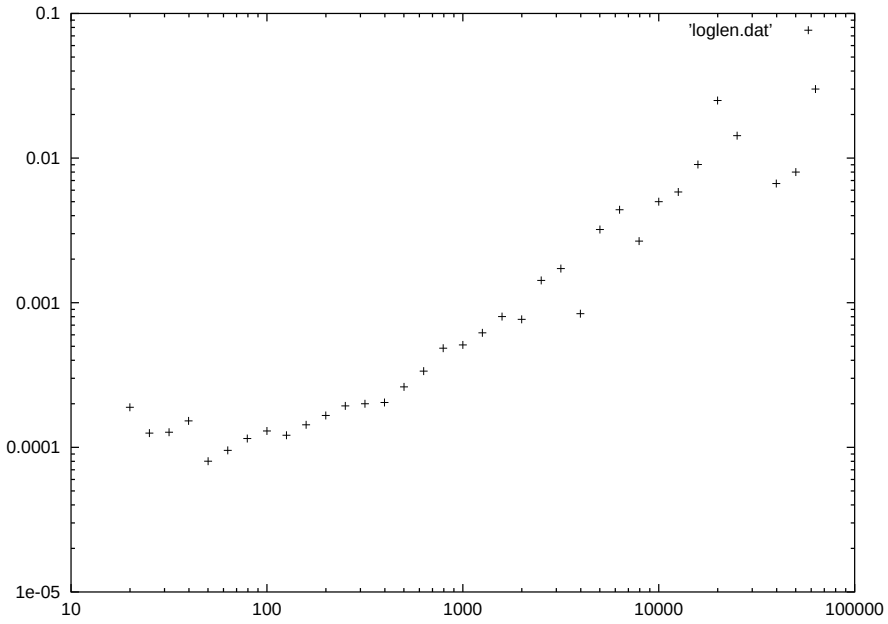

Figure 2: $P(D|l)$ for the TREC8 Ad Hoc collection

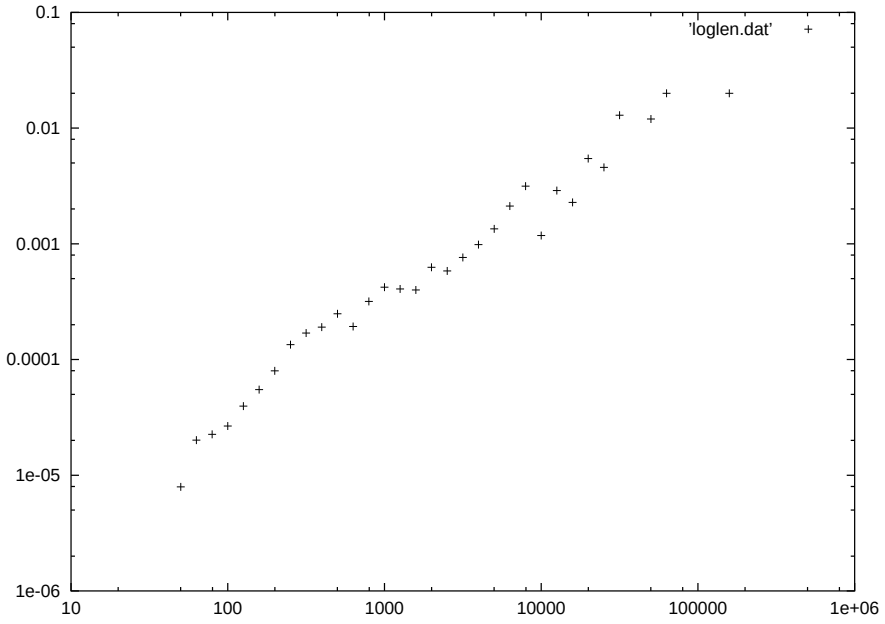

Figure 3: $P(D|l)$ for the TREC8 small web collection

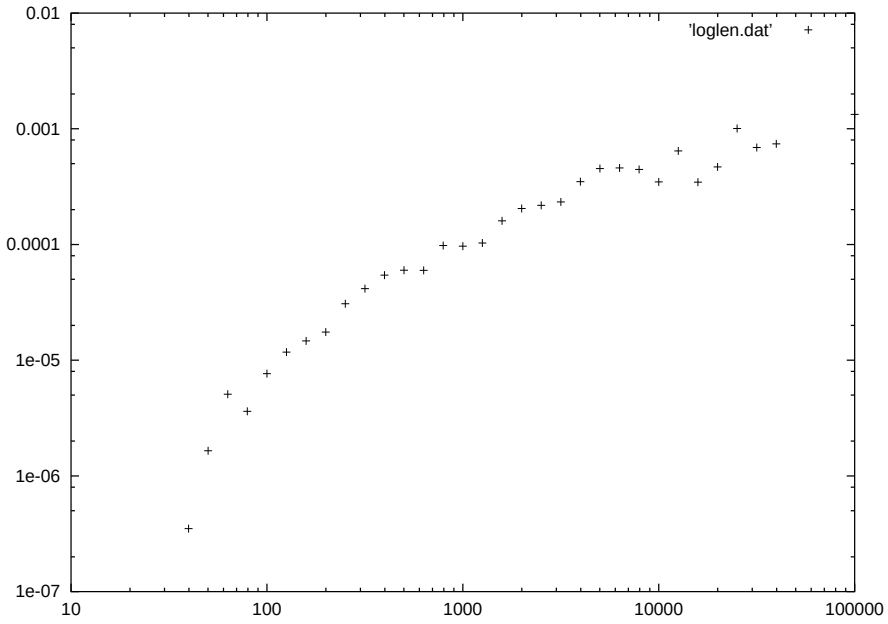

Figure 4: $P(D|l)$ for the TREC9 main web collection

the qrel file, but we took bins on a log scale. Subsequently, we computed

$$P(D_k\text{isRel}|\text{dlen}_k \in \text{bin}_k) = \frac{P(\text{dlen}_k \in \text{bin}_k|D_k\text{isRel}).P(D_k\text{isRel})}{P(\text{dlen}_k \in \text{bin}_k)} \tag{7}$$

The plots for the web collections seem quite comparable and distinct from the adhoc plots, which in turn are also quite comparable. Especially shorter Ad Hoc documents are relatively much more relevant than their web counterparts. This could be explained by the fact that shorter web documents are often just placeholders for links or pictures. We might be able to improve the performance of our runs by taking this fact into account while estimating document priors.

4 Exploitation of links

We used different link-based techniques to recalculate the scores for the top 1000 documents retrieved by a title-only, content-only run (tnout9t2). Although in last year's TREC, adding link information didn't seem to help, we hope that the higher density of links in this year's collection can improve the results. Below, we first discuss the different approaches and then compare the results to the original content-only run.

We used two different approaches. The first one is the well-known Kleinberg algorithm of hubs and authorities [2]. We took the top N documents with their in and out links and computed hubs and authoroties on that set. We then normalised the content only scores in the same way the scores are normalised in the kleinberg algorithm (equation 8). The normalised scores and kleinberg scores are then summed.

$$newscore(d) = \frac{score(d)}{\sum_{d \in docs}(score(d))^2} \tag{8}$$

The second approach is based on co-citation [5] and bibliographic coupling [1]. The assumptions behind the use of these measures to adjust the document scores are the following. If the set of documents that document A refers to is similar to the set of documents that document B refers to, then document A and B are similar. If the set of documents

that refer to A is similar to the set of documents that refer to B, then A and B are similar. First we analysed last year's results. We propagated the relevance judgements along the links and computed the following scores:

$$Inlinkrel(d) = \frac{\sum_{i \in inlinks(d)} relevancy(i)}{\#inlinks(d)} \tag{9}$$

$$Outlinkrel(d) = \frac{\sum_{i \in outlinks(d)} relevancy(i)}{\#outlinks(d)} \tag{10}$$

$$Citerel(d) = \frac{\sum_{i \in inlinks(d)} Outlinkrel(i)}{\#outlinks(d)} \tag{11}$$

$$Bibcouplrel(d) = \frac{\sum_{i \in outlinks(d)} Inlinkrel(i)}{\#inlinks(d)} \tag{12}$$

In table 2 the average scores are shown for the whole collection, the assessment pool and the relevant set. Relevant documents have higher cocitation and bibliographic coupling scores than an average (judged) document. We used this information to recalculate the scores of a topic only run in the following way. We took the top N retrieved documents and propagated their scores allong their in and outlinks calculating cocitation and bibcoupling scores analogously to the cocitation and bibliographic coupling relevancies in equations 11 and 12. We used the resulting cocitation and bibliographic coupling scores to weigh the original content-only scores.

	Relevance	Inlinkrel	Outlinkrel	Citerel	Bibcouplrel
Collection (WT2g)	0.00921076	0.012829736	0.004616373	0.00728053	0.006673368
Assessment pool	0.050987634	0.010024281	0.004668967	0.010140689	0.010697012
Relevant set	1	0.064735196	0.026876365	0.126311025	0.137629731

Table 2: Average indirect relevancy

Due to some misunderstanding about the calculation of the scores, in the official content-link runs the content-only scores are reweighed by multiplying the content-only scores and the link scores (i.e. cocitation scores). However, the original content-only scores are composed of a sum of logarithms of different weights. In unofficial runs (with runtags ending in log), the link scores are properly combined with the content only scores. The results for the different runs can be seen in table 3

Adding link information decreases or at the best doesn't influence the average precision. When we take a closer look at the different link runs and compare them to the content only title run, we see that most runs hardly differ from it. The only run that differs a lot is tnout9t2lk50. In this run, the authority scores are added to the normalised

runtag	official run	description	average precision
tnout9t2	yes	title only content run	0.1801
tnout9t2lk50	yes	kleinberg on top 50	0.0488
tnout9t2lc10	yes	cocitation on top 10	0.1630
tnout9t2lc50	yes	cocitation on top 50	0.1337
tnout9t2.klein50log	no	kleinberg on top 50	0.1803
tnout9t2.coc10log	no	cocitation on top 10	0.1786
tnout9t2.coc50log	no	cocitation on top 50	0.1784
tnout9t2.bib10log	no	bibcoupling on top 10	0.1691
tnout9t2.bib50log	no	bibcoupling on top 50	0.1642

Table 3: Content-Link results

content only scores. This means that the link information is regarded equally important as the content information. In all the other runs, link information was only used to reweigh the content information. When we use kleinberg authority scores for reweighing (tnout2.klein50log), this doesn't change the results of the content run . Even though this year's collection is bigger and has more links than last year's collection, the use of link information does not seem to improve the retrieval results. One of the reasons for this might be that TREC topics are not suitable for using link information because they are too specific. This year's TREC topics on average have only 47.4 relevant documents, so the changes of being many links between them are rather small. Another reason is that there's a lot of garbage in the link information. The basic assumption behind these link based methods is that pages are topically related if they are linked to each other. Obviously, this isn't necessarily the case. Many links on the web refer to creators of the page, sponsors, friends or other pages without a topical relation to the source page of the link. Classifying links in advance into meaningful and meaningless links on the basis of for example the anchor text might help.

5 Conclusion

To our surprise, the web task turned out to be more difficult than we expected. Firstly, web documents (even though the collection has been cleaned) contain a lot of trash, often in the form of of incorrect HTML. The HTML parsing component had to be adjusted to be able to handle this kind of material. Secondly, web documents contain a lot of misspellings. These are often very rare terms. When such terms occur in a pseudo feedback document, they will have a bad influence on the pseudo feedback process, because these terms will receive a high weight. Finally, the title only queries posed a problem. Some queries contained typos, involving digits (topic 487). Our engine simply discarded those terms. The tokenizer has to be updated as well to deal with years. Four-digit years were important query concepts in quite a few queries, they were discarded by our term extraction module.

The content-only runs were finally run with some small variants of our standard IR model. Both the full and title runs perform well (33 resp 37 topics) above median, confirming the adequacy of the model. The runs which additionally analyzed link information in order to rescore the runs, were not able to improve on average precision. This confirms the general result of the TREC8 web runs. We hope to improve the link analysis in the future by looking at the anchor texts.

References

[1] M.M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25, 1963.

[2] J.M. Kleinberg. Authorative sources in a hyperlinked environment. In *Proceedings of 9th ACM-SIAM Symposium on Discrete Algorithms*, pages 668–377, 1998.

[3] W. Kraaij, R. Pohlmann, and D. Hiemstra. Twenty-one at TREC-8: using language technology for information retrieval. In *The Eighth Text Retrieval Conference (TREC-8)*. National Institute for Standards and Technology, 2000.

[4] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–29, 1996.

[5] H. Small. Co-citation in the scientific literature: A new measure of the relationship between documents. *J. American Soc. Info. Sci.*, 24:265–269, 1973.