High Precision Retrieval Using Relevance-Flow Graph

Jangwon Seo
Center for Intelligence Information Retrieval
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA 01003
jangwon@cs.umass.edu

Jiwoon Jeon
Google, Inc.
Mountain View, CA 94043
jjeon@google.com

ABSTRACT

Traditional bag-of-words information retrieval models use aggregated term statistics to measure the relevance of documents, making it difficult to detect non-relevant documents that contain many query terms by chance or in the wrong context. In-depth document analysis is needed to filter out these deceptive documents. In this paper, we hypothesize that truly relevant documents have relevant sentences in predictable patterns. Our experimental results show that we can successfully identify and exploit these patterns to significantly improve retrieval precision at top ranks.

Categories and Subject Descriptors

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

General Terms

Algorithms, Measurement, Experimentation

Keywords

relevant sentence, relevance flow, re-ranking

1. INTRODUCTION

To achieve high precision retrieval, filtering out highly ranked non-relevant documents is crucial. These documents look relevant in traditional bag-of-words models because they contain many query terms like relevant documents. In this paper, we show these deceptive, non-relevant documents can be successfully identified and demoted by analyzing the change of relevance scores at the sentence level. We especially focus on the strength and the position of the relevant sentences in the document.

To analyze the spatial distribution of relevant sentences in a document, we plot relevance scores of sentences versus their locations in the document as shown in Figure 1. We call this graph a *relevance-flow graph*. The locations of peaks show positions of relevant sentences, and the height of peaks represent the strength of relevance. From the graph, we extract a set of features that can capture various aspect of the graph. Using training data sampled from the TREC collection we learn a probabilistic model that can distinguish relevance-flow patterns from relevant documents. Top documents returned by the baseline search engine are re-ranked solely using the classifier scores. Our evaluation result shows that this approach significantly improves precision, especially at top ranks.

There have been efforts to analyze the distributional patterns of query terms to measure term proximity scores [4]. However, as far as we know, there has been little attempt to infer document relevance based on sentence-level relevance analysis. In this paper, we show this kind of approach is promising.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

2. RELEVANCE-FLOW GRAPH

To measure relevance scores of sentences, we use normalized query likelihood scores as estimates. The query likelihood score is the probability of query Q given the Dirichlet-smoothed unigram language model of a sentence S, i.e. $P(Q|S)$ [1]. For a query, we compute the query likelihood scores for all sentences in the top N documents returned from the baseline search engine. Then, each score is normalized by,

$score_{normalized}=(score–score_{min})/(score_{max}–score_{min})$ where $score_{max}$ and $score_{min}$ are the maximum and the minimum relevance scores across the top N documents. We call the normalized query likelihood score the *relevance level*. If the relevance level of a sentence is greater than 0.5, then we call the sentence a *peak*. A sentence at the peak can be considered as an estimate for a relevant sentence. In addition, positions of sentences are also normalized for comparison across different documents: 0 for the first sentence and 1 for the last sentence.

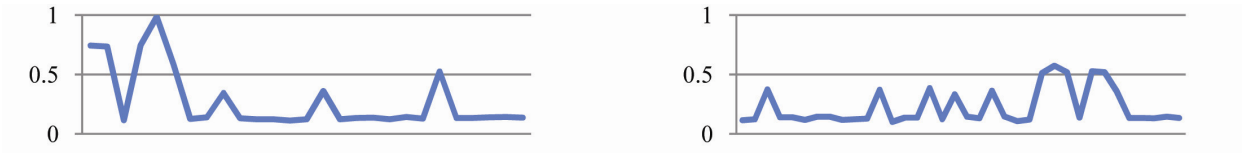
The relevance-flow graph visually shows the fluctuation of relevance level inside of the document. This graph is often useful to understand why one document is more relevant than the other. For example, Figure 1 shows relevance-flow graphs of a relevant document (Figure 1(a)) and a non-relevant document (Figure 1(b)) for a given query. Both documents have similar log likelihood scores from the baseline search engine. Figure 1(a) has an early peak followed by a few smaller peaks while Figure 1(b) has many smaller peaks at the end of the document. Intuitively, having an early peak is a good sign because many writers put key sentences at the beginning of their articles [5]. A high peak can be also considered as a positive sign because high peaks mean that the majority of query terms appear in the sentence, that is, the proximity among query terms is well-preserved. Our retrieval system can successfully identify these differences and demote the non-relevant document in Figure 1(b). Exploiting these observations is impossible in traditional bag-of-words retrieval models.

3. INFERRING DOCUMENT RELEVANCE

We learn a statistical model which is able to predict the relevance of a document from the relevance-flow graph of the document. The logistic regression model is used with the following six features extracted from the graphs.

Mean and Variance of Relevance Level (F1.1 and F1.2)

The arithmetic mean of relevance levels shows how relevant a document is at large. This can also be interpreted as COMBAVG in rank fusion [2]. High variance values imply many peaks and valleys in the graph.



(a) Query:119, Document: AP891108-0106 (Relevant) (b) Query: 119, Document: AP900914-0164 (Non-Relevant)

Figure 1. Relevance-flow graph examples.

Peak to Sentence Ratio (F2)

To investigate relations between the number of peaks and relevance, we use (*#peak* / *#sentence*) as a feature.

The first peak position (F3)

If a document is relevant, then relevant sentences often appear in the beginning of the document. Troy et al. exploited this property and reported improved retrieval performance [5]. Xue and Zhou used this property for text classification tasks [6]. We use the position of the first peak in a document as a feature.

Mean and Variance of peak positions (F4.1 and F4.2)

The mean of peak positions roughly shows where relevant sentences appear. The variance shows how relevant sentences are spatially distributed in a document.

Table 1. Feature Weights

F1.1	F1.2	F2	F3	F4.1	F4.2
0.68	14.03	-2.04	-1.64	0.92	-0.88

Table 1 shows learned weights for the features. Not surprisingly, the mean of relevance levels (F1.1) has a small weight. Since all documents in the initial search result are competitive under the bag-of-words model, the relevance level itself is no longer discriminative. On the other hand, the variance of the relevance levels (F1.2) has the largest weight. This shows that relevant documents are likely to contain a few highly relevant sentences rather than many medium or low relevance level sentences. The peak to sentence ratio (F2) also shows a similar aspect. F2 has a negative relation to relevance. That is, fewer peaks are preferred. This may sound strange. However, because we know that most top ranked documents have similar query likelihood scores, this can be interpreted as a few "high" peaks are preferred to many "low" peaks. Therefore, the larger weights of F1.2 and F2 imply importance of term proximity. There is a negative relationship between the first peak position (F3) and relevance. In other words, the early appearance of a relevant sentence is preferred as expected. Both the mean (F4.1) and the variance (F4.2) of the peak positions have little impact and removing these features did not hurt our re-ranking performance. Overall, F1.2 is the most important feature. F2 and F3 are also useful.

4. EXPERIMENT AND RESULTS

For evaluation, we used the AP collection and title queries of topics 51-200 in the TREC corpora. We split the queries by query-id *mod* 3, i.e. queries whose id *mod* 3 = 0 or 1 into a training set (100 queries) and the others into a test set (50 queries). As preparation for plotting relevance-flow graphs, each document was segmented into sentences using the MXTerminator [3].

We retrieved the top 15 documents for each query using the bag-of-words model implemented in the Indri[1] search engine, where the unigram language model was used with the Dirichlet smoothing parameter μ_{doc}=3600, which produced the best retrieval performance. Unigram "sentence" language models for relevance-flow graphs are smoothed with the collection language model (Dirichlet smoothing parameter $\mu_{sentence}$=300).

Table 2. Retrieval performance. A superscript * indicates a statistically significant improvement on the initial result. (sign test with *p*-value < 0.05)

	P@1	P@5
Initial result	0.380	0.336
Re-ranked result	0.480 (+26%)*	0.396 (+18%)*

We re-ranked the initial search result according to predicted relevance. Since our purpose is to achieve high precision in the top results, we use precision at 1 (P@1) and precision at 5 (P@5) as evaluation metrics. Table 2 shows the retrieval performance. The re-ranked results show statistically significant improvements over the initial result for both metrics.

5. CONCLUSION AND FUTURE WORK

In this paper, we demonstrated that analyzing spatial distribution patterns of relevant sentences has major potential to improve precision of retrieval systems. This work is just the first step of our efforts to understand document relevance via relevance analysis at the sentence level. We plan to explore different methods to create relevance-flow graphs and more descriptive features that can capture various aspects from the graphs. Finally, we will investigate new scoring functions which can seamlessly combine traditional document-level scores with scores based on our sentence-level analysis.

6. ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval (CIIR) and in part by NSF grant #IIS-0711348. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] W. B. Croft and J. Lafferty. *Language Modeling for Information Retrieval*. Kluwer Academic Publishers, Norwell, MA, USA, 2003.

[2] E. Fox and J. Shaw. Combination of multiple searches. In *Proc. of TREC-2*, 1994.

[3] J. Reynar and A. Ratnaparkhi. A maximum entropy approach to identifying sentence boundaries. In *Proc. of ANLP*, 1997.

[4] T. Tao and C. Zhai. An exploration of proximity measures in information retrieval. In *Proc. of SIGIR '07*, pages 295–302, 2007.

[5] A. D. Troy and G. Zhang. Enhancing relevance scoring with chronological term rank. In *Proc. of SIGIR '07*, pages 599–606, 2007.

[6] X.-B. Xue and Z.-H. Zhou. Distributional features for text categorization. In *Proc. of ECML '06*, pages 497-508, 2006.

[1] http://www.lemurproject.org/indri/