Selecting Expansion Terms in Automatic Query Expansion

Hiroko Mano
Software Research Center
Ricoh Company, Ltd.
1-1-17 Koishikawa, Bunkyo-ku
Tokyo 112-0002 JAPAN
mano@src.ricoh.co.jp

Yasushi Ogawa
Software Research Center
Ricoh Company, Ltd.
1-1-17 Koishikawa, Bunkyo-ku
Tokyo 112-0002 JAPAN
yogawa@src.ricoh.co.jp

1. INTRODUCTION

In automatic query expansion, where queries are automatically expanded with terms not in the original query but extracted from initially retrieved top-ranked documents, each term in the top-ranked documents is evaluated for its usefulness as an expansion term to be added to the original query. As the evaluation measure, we have used since [1] a Term Selection Value function based on a combination of a relevance weight [3] derived from document frequencies and a sum of within-document term frequencies, so that terms that are specific to top-ranked documents (relevance weight factor) and representative of each top-ranked document (term frequency factor) would be rated highly.

While query expansion based on the TSV function above contributed to improvement in retrieval effectiveness in the past experiments, recent experiments suggested selecting expansion terms based on both the relevance weight and the term frequencies might not always be the best strategy; specifically, when the top-ranked documents, *assumed* to be relevant to the topic, turn out to be mostly non-relevant, the relevance weight seemed to lead to inappropriate selection of expansion terms.

In this paper, we investigate how a relevance weight affects expansion term selection as the number of relevant documents in the top-ranked documents decreases and examine the effectiveness of an alternative approach of not using a relevance weight in expansion term evaluation.

2. EXPERIMENT

Obviously, the quality of query expansion may well be influenced by the quality of the initially retrieved top-ranked documents, in particular, how many of them are actually relevant; the question is whether the degree of influence varies depending on the use of relevance weights when expansion terms are selected. To investigate this, we conducted the following experiment using the TREC-8 Web Track topics and its WT2g document collection and the TREC-9 Web Track topics and its WT10g document collection.

SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA.

2.1 Term Selection Value functions

We compared two term selection measures to select expansion terms: one that uses a relevance weight (TSV-1), and the other that does not use a relevance weight (TSV-2), i.e.,

$$TSV\text{-}1 = w_t \cdot Sum_t,$$
$$TSV\text{-}2 = Sum_t,$$

where w_t is a relevance weight and Sum_t is a sum of within-document term frequencies. The functions are variations of Term Selection Value in [4].

The relevance weight w_t is the same as the relevance weight our retrieval engine [2], based on the probabilistic model, assigns to both original and expansion terms after top-ranked documents are identified in initial retrieval and follows a weighting scheme similar to [3]:

$$\begin{aligned} w_t = & \frac{k_5}{k_5+\sqrt{R}} \log\left(k_4' \frac{N}{N-n_t} + \frac{n_t}{N-n_t}\right) \\ & + \frac{\sqrt{R}}{k_5+\sqrt{R}} \log \frac{r_t+0.5}{R-r_t+0.5} \\ & - \frac{k_6}{k_6+\sqrt{S}} \log \frac{n_t}{N-n_t} \\ & - \frac{\sqrt{S}}{k_6+\sqrt{S}} \log \frac{s_t+0.5}{S-s_t+0.5}, \end{aligned}$$

where R is the number of relevant documents, r_t is the number of relevant documents containing the term, S is the number of non-relevant documents, s_t is the number of non-relevant documents containing the term, N is the number of documents in the collection, n_t is the number of documents containing the term, and k_4', k_5 and k_6 are parameters.

The sum of within-document term frequencies Sum_t, included in both TSV functions, uses the within-document frequency normalized by the document length, i.e.,

$$Sum_t = \left(\frac{\sum_{d \in R} \frac{f_{t,d}}{K+f_{t,d}}}{R} - \beta \frac{\sum_{d \in S} \frac{f_{t,d}}{K+f_{t,d}}}{S} \right),$$
$$K = k_1 \left((1-b) + b \frac{l_d}{l_{ave}} \right),$$

where $f_{t,d}$ is the within-document frequency of the term, l_d is the document length, l_{ave} is the average document length, and k_1, b and β are parameters.

Parameter values were determined after preliminary experiments to find a reasonable combination. (S was set to 0 in our experiment.) There was also a cut-off measure applied to limit terms that appear in a few top documents.

Using the above measures, up to 30 terms were selected as expansion terms per topic.

2.2 Top-ranked document sets

We created three conditions of top-ranked documents, two *simulated* and one real: worst-possible, best-possible and realistic. To simulate the worst case scenario where initial retrieval fails to turn up any relevant document, a set of non-relevant documents (NONE set) was constructed by weeding out all relevant documents from top-ranked documents that contain query terms while retaining the relative rank of each non-relevant document that will remain in the set. Similarly, a set of relevant documents (ALL set) was created by picking up only relevant documents from the top-ranked documents.

For comparison purposes, a set of top-ranked documents as they were obtained in initial retrieval without modification (SOME set) was also prepared. In the TREC-8 runs, approximately 48 to 50% of the set were relevant documents on the average. In the TREC-9 runs, the ratios of relevant documents were in the range of approximately 25 to 35%. There were at most 10 documents in total in each of the three sets.

3. RESULT

The experimental runs resulted in the average precision measurements in Tables 1 and 2. Queries were created using either the title field only or the title and desc fields.

The result shows that, with the ALL set, selecting terms using TSV-1 resulted in the higher average precision than selecting terms using TSV-2. On the other hand, with the NONE set, using TSV-1 affected retrieval more adversely than using TSV-2. The outcomes using the SOME set also suggest that the more the top-ranked document set contains relevant documents, the greater TSV-1 outperformed TSV-2. Note also that using original queries with no expansion produced better results in all runs with the NONE set, but with the SOME set, only in the runs using TSV-1 in TREC-9 runs. This implies that there is at least one form of query expansion that is effective when around one third or so of top-ranked documents is relevant.

When we compare the terms selected, those selected by TSV-1 resulted in the average document frequency considerably lower than that of those selected by TSV-2, regardless of which set the terms came from. This indicates that the TSV-1 tends to consistently favor less common terms over common terms, despite the fact that the function incorporates within-document term frequencies.

Another observation about the selected terms is that, with TSV-2, the average document frequency of the selected terms increases as the number of non-relevant documents in the top-ranked documents increases. We found this particularly interesting since it seems to serve as a mechanism to counter potentially negative impact of selecting terms based on non-relevant documents. This tendency may indicate that as non-relevant documents dominate the top-ranked documents, it becomes more difficult to select terms that are commonly shared in the top-ranked documents unless they are high document frequency terms.

Table 1: TREC-8 topics and data collection

		title only	title + desc
ALL	TSV-1	0.4649	0.4707
	TSV-2	0.4314	0.4374
SOME	TSV-1	0.3503	0.3687
	TSV-2	0.3523	0.3693
NONE	TSV-1	0.2848	0.3091
	TSV-2	0.2968	0.3184
No expansion		0.3247	0.3420

Table 2: TREC-9 topics and data collection

		title only	title + desc
ALL	TSV-1	0.3474	0.4074
	TSV-2	0.3122	0.3725
SOME	TSV-1	0.2021	0.2427
	TSV-2	0.2150	0.2705
NONE	TSV-1	0.1686	0.1866
	TSV-2	0.1810	0.2182
No expansion		0.2073	0.2608

4. CONCLUSIONS

The experiment showed that term selection using a relevance weight was sensitive to how many of the top-ranked documents were actually relevant and that, with this method of term selection, the range of fluctuation resulting from the difference in the ratio of relevant documents was quite wide. The result also suggests that, under many realistic circumstances, dropping relevance weights in term selection seems to produce more stable results. We also found it useful to consider best case and worst case scenarios to gain insight of what works under what circumstances.

5. REFERENCES

[1] Y. Ogawa, H. Mano, M. Narita, and S. Honma. Structuring and expanding queries in the probabilistic model. In *The Eighth Text REtrieval Conference (TREC-8)*, pages 541–548. National Institute of Standards and Technology, November 2000.

[2] Y. Ogawa, H. Mano, M. Narita, and S. Honma. Structuring and expanding queries in the probabilistic model. In *The Ninth Text REtrieval Conference (TREC-9)*, to be published. National Institute of Standards and Technology, November 2001.

[3] S. Robertson and S. Walker. On relevance weights with little relevance information. In *Proceedings of 20th ACM SIGIR Conference*, pages 16–24. Association for Computing Machinery, July 1997.

[4] S. Walker, S. Robertson et al. Okapi at TREC-6: Automatic ad hoc, VLC, routing, filtering and QSDR. In *The Sixth Text REtrieval Conference (TREC-6)*, pages 125–136. National Institute of Standards and Technology, November 1997.