Investigation of Partial Query Proximity in Web Search

Jing Bai, Yi Chang, Hang Cui, Zhaohui Zheng, Gordon Sun, Xin Li
Yahoo! Inc., 701 First Avenue, Sunnyvale, CA
94089 USA
{jingbai, yichang, cui, zhaohui, gzsun, xinli}@yahoo-inc.com

ABSTRACT

Proximity of query terms in a document is an important criterion in IR. However, no investigation has been made to determine the most useful term sequences for which proximity should be considered. In this study, we test the effectiveness of using proximity of partial term sequences (n-grams) for Web search. We observe that the proximity of sequences of 3 to 5 terms is most effective for long queries, while shorter or longer sequences appear less useful. This suggests that combinations of 3 to 5 terms can best capture the intention in user queries. In addition, we also experiment with weighing the importance of query sub-sequences using query log frequencies. Our preliminary tests show promising empirical results.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval] - *Query formulation*

General Terms

Algorithms, Measurement, Experimentation

Keywords

Term proximity, Information retrieval

1. INTRODUCTION

In addition to the traditional criteria concerning individual query terms such as *tf* and *idf*, the proximity between query terms in a document is often believed to be a useful criterion for document ranking [3,4]. The rationale is that documents in which the query terms appear more closely have a better chance to represent the concept the query implies. Several approaches using term proximity have been suggested in recent IR studies. For example, Tao and Zhai [4] proposed to reward a document according to either the proximity of all the query terms together or the proximity of each pair of terms. The first method requires all the query terms to appear together in a text window or text span. This requirement may be reasonable for short queries (with fewer than 3 terms), but could be too strong for longer queries. Indeed, it is often the case that not all terms in a query such as "second hand car for sale in California" appear in the same text span in relevant documents. Query terms may scatter in disjoint text spans. On the other hand, considering only word pairs may not be sufficient for long queries. For the above example, if we only consider proximity of word pairs such as "second car" or "hand car", the underlying concept "second hand car" cannot be correctly covered. As such, longer term sequences (in this case, of 3 terms) are needed.

It is important to deal with long queries in Web search, as a non negligible part of Web search queries are long queries containing 4 or more words: a sampling of 100K queries on a commercial search engine shows that around 20% of the unique queries contain 4 or more words. For these queries, the approaches proposed in [4] represent the two extreme cases, which enforce either only the adjacent two term proximities or the proximity of the all terms. In this paper, we show that both approaches fail to deal with long queries correctly. The proximity of partial combinations of 3 to 5 terms is more appropriate for long queries.

Metzler and Croft [3] have also investigated the utilization of term proximity in a language modeling setting. They showed that by integrating term proximity, improvements in IR effectiveness can be obtained on TREC collections. However, no investigation has been made directly on Web search relevance so far. The approach proposed by [3] also investigates the combinations of all lengths. The results of our experiments show that it may not be necessary to extend the combinations beyond 5 terms.

The primary goal of this study is to see how term combinations (n-grams) of different lengths affect IR effectiveness (measured in Discounted Cumulative Gain – DCG [2]). Our experiments show that partial combinations of 3 to 5 terms should be considered for proximity and integrated into the ranking function.

Term sequences in queries may not be equally representative or meaningful. Another important question is how one can determine meaningful sequences and assign them a higher importance for ranking. In this study, we propose to exploit query logs for this purpose: more frequent term sequences are considered to be more meaningful. Our preliminary experiments show this is helpful.

2. USING PROXIMITY OF TERMS

Without losing generality, in our study, we assume that a web search ranking function could be represented by two parts of information: one is a proximity-independent ranking function $f_1(Q,D)$ for a query Q and a document D, while the other is a proximity-related function $f_2(Q,D)$. These functions can be combined in different ways to produce a final ranking function $f(Q,D)$. For example, both [3] and [4] use $f(Q,D) = f_1(Q,D) + f_2(Q,D)$, in which the two component ranking functions are used at equal importance. In practice, one criterion may be more important than another. Therefore, we will use weighted sum as follows:

$$f(Q,D) = f_1(Q,D) + w^*f_2(Q,D)$$

The relative weight w is to be tuned using training data. In this study, $f_1(Q,D)$ is an existing ranking function without term proximity, which contain a set of parameters to be tuned. These parameters and the weight w are tuned together using Gradient Boosting Tree [1]. The proximity-related criterion $f_2(Q,D)$ is defined on n-grams of different lengths as follows:

$$f_2(Q,D) = \sum_{T_i \in Q} \sum_{L_j \in D} prox(T_i, L_j)$$

where T_i is an n-gram (with a predetermined n) in Q and L_j is a

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

line or span in D with a length limit (16 in our experiments). $prox(T_i, L_j)$ is defined as follows:

$$prox(T_i, L_j) = 1/\exp(miss * \alpha + extra * \beta)$$

where α and β are two parameters fixed in (0,1) to penalize L_j if a term in T_i is missing in it, in bad order or if it contains an extra term between the terms in T_i. *miss* and *extra* are the numbers of missing and extra terms.

We first test the impact of each type of n-gram separately. The test is performed on a set of 2,669 queries, among them, 579 are long queries (4 or more terms). These queries are extracted from Yahoo! search engine. There are 80K query-URL pairs in total. These pairs are manually judged by subjects. We use a 5-fold cross-validation in the test: one part of the long queries (together with the corresponding URLs) is used for test while 4 other parts (including short queries) are used to train the parameters, in particular, the relative weights w_1 and w_2.

As most users on Web search are interested in top results only, we use DCG1 and DCG5 [2] as our measures of retrieval effectiveness. The following table shows the impact of each n-gram on DCG1 and DCG5. The baseline method only uses $f_1(Q, D)$ for ranking.

Table 1. Effect of integrating different n-grams

	baseline	2-gram	3-gram	4-gram	5-gram	6-gram
DCG1	1.94	1.97 +1.4%	2.07 +6.8%*	2.13 +9.7%*	2.10 +7.9%*	2.05 +5.5%*
DCG5	4.63	4.56 -1.5%	4.70 +1.5%*	4.71 +1.8%	4.72 +2.0%*	4.60 -0.6%

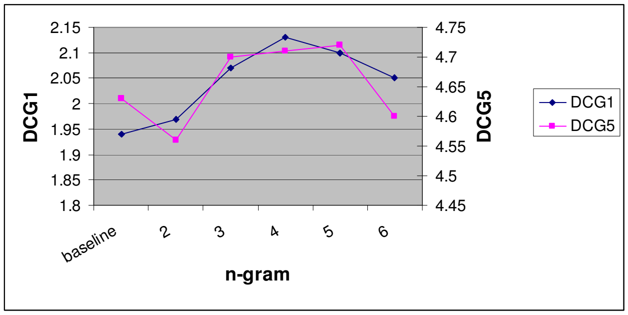

Figure 1. DCG1 and DCG5 with different n-grams

We can see that bigrams can improve slightly DCG1 for long queries, but it decreases DCG5. This shows that proximity of bigrams has a limited impact on IR effectiveness. This conclusion is similar to that of [4] when the average distance of word pairs is used in their approach.

When longer n-grams are used, we observe larger improvements in DCG1 and all the improvements are statistically significant at the level of $p<0.05$ (marked with *). For DCG5, we observe the improvements with 3- to 5-grams and those with 3-grams and 5-grams are statistically significant.

The above experiments indicate that the consideration of term proximity of different size can all improve the top-1 retrieval result. However, only the sequences of middle range sizes (3-5) can improve the top-5 results. This shows that it is not very useful to exploit proximity of sequences longer than 5 terms in practice.

A possible explanation of the above observation is that many concepts used in user queries correspond to 3 to 5 words. Shorter sequences are unable to cover these concepts correctly, while longer sequences often aggregate additional terms to the concepts, making them less meaningful. This gives rise to the question of how we can select and weight meaningful term sequences. Ideally, we should consider the proximity of meaningful term sequences only. In the next section, we propose to use query logs to determine the meaningfulness of a term sequence.

3. ESTIMATING THE IMPORTANCE OF N-GRAMS USING QUERY LOGS

Our assumption here is that the more a term sequence appears in many queries, the more the sequence is meaningful. Each term sequence can then be weighted according to their frequency in query logs. In our experiments, we use query logs of 6 months on Yahoo! search engine. Let $freq(T_i)$ be the frequency of n-gram in query logs. The weigh for an n-gram T_i in a query is estimated by $w(T_i) = Z^{-1}\log(freq(T_i))$, where Z is normalization factor over the whole query. The new function $f_2'(Q, D)$ is defined as follows:

$$f_2'(Q, D) = \sum_{T_i \in Q} \sum_{L_j \in D} w(T_i) \times prox(T_i, L_j)$$

We have only tested this approach for bigrams. For longer n-grams, we have more problem of data sparseness. Larger query log data is required. With the above ranking function for 2-grams, we obtain 2.08 and 4.64 in DCG1 and DCG5 respectively, or 7.24% and 0.19% improvements compared to the baseline method. Compared to the case of unweighted bigrams shown in Table 1, DCG1 and DCG5 are increased. This preliminary result suggests that query logs can help determine the most important term sequences for which term proximity should be enhanced. In our future study, we will perform experiments for longer n-grams using larger query logs.

4. CONCLUSION

Proximity of query terms in documents is an important factor for ranking, but no investigation has been made on the usefulness of proximity of different term sequences. In this study, we investigated the impact of proximity measure of 2- to 6-grams. Our experiments on real Web search data showed that the most useful sequences for long queries are those of 3 to 5 words.

We also tested the utilization of query logs to determine the meaningfulness of a term sequence. Our preliminary result showed that query logs can provide useful information for this. The resulting ranking function can be better than when all term sequences are weighted equally.

5. REFERENCES

[1] Friedman, J.H., Greedy function approximation: A gradient boosting machine, 1999, http://www.salford-systems.com/doc/GreedyFuncApproxSS.pdf.

[2] Järvelin, K. & Kekäläinen, J. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20, 422-446, 2002.

[3] Metzler, D. and Croft, W.B., A Markov Random Field Model for Term Dependencies, *SIGIR 2005*, pp. 472-479.

[4] Tao T., Zhai, C., An exploration of proximity measures in information retrieval. *SIGIR 2007*, pp. 295-302.