Higher Precision for Two-Word Queries

K. L. Kwok
C. S. Dept., Queens College, CUNY,
Flushing, NY 11367
Tel: +1 718 997 3482
kwok@ir.cs.qc.edu

ABSTRACT
Queries have specific properties, and may need individualized methods and parameters to optimize retrieval. Length is one property. We look at how two-word queries may attain higher precision by re-ranking using word co-occurrence evidence in retrieved documents. Co-occurrence within document context is not sufficient, but window context including sentence context evidence can provide precision improvements at low recall region of 4 to 10% using initial retrieval results, and positively affects pseudo-relevance feedback.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval.

General Terms: Experimentation.

Keywords: Term co-occurrence; re-ranking of retrieval list.

1. INTRODUCTION
2-word queries (after stop-word removal) present a unique opportunity to study how word co-occurrence may influence retrieval effectiveness. In such a query, the two words can form a phrase (such as noun-noun) that has more specific meaning than the single words alone. This phrase is either in the order presented, or the reverse if a preposition (e.g. 'of') exists between them. If they do not form a phrase, we assume they form a list and that both words are important because the query is so short. In contrast, 1-word queries present no opportunity to use relationships between words. 3 (or more)-word queries lead to ambiguity as to which 2-word combinations are more meaningful (in addition to the 3-word construct). 2-word queries often have low effectiveness (usual with short queries) unless the words are very specific. We study how co-occurrence of the two query words may be used to re-rank the retrieval based on single stems to achieve higher precision. Higher precision at 10 to 30 documents retrieved could also be important for pseudo-relevance feedback (PRF). Previous studies of re-ranking [1,2] normally consider a query set of all types, and use a single technique without discrimination. We believe each query may have its own specific properties, especially ultra-short ones, and a more individualized approach using varied techniques or parameters for different query types may be a better strategy. One property is length, and we start with query length of two.

SIGIR'02, August 11-15, 2002, Tampere, Finland.
ACM 1-58113-561-0/02/0008.

2. METHODOLOGY
The different ways in which query words appear in documents may provide evidence of relevance. [3,4] employed such occurrence patterns for all query types to do primary ranking of documents. Here, we use these patterns to re-rank the top n documents of retrieval based on single stems. One common co-occurrence factor to study is the context unit in which query terms occur: document, sentence, or a window within sentence. For document context (coordinate matching) we followed [1] to re-rank all 2-word matching before single word matching, and retain the original RSV's (retrieval status value) to resolve ties. One may also use sentence context to provide tighter control of the matching to avoid erroneous evidence. In coordinate matching, binary count was used. Original retrievals were performed using Porter's stemming. In [1], it has also been reported that for re-ranking with un-stemmed words provide better evidence than stemmed words.

For window co-occurrence, we include more detailed considerations such as: 1) weight modification for re-ranking; 2) matching order and window size w; 3) number of co-occurrence matches in a document; and 4) hardness of improvement. Previous works in phrase retrieval have tried to define the weight of a phrase match [5], which is then added to the normal RSV due to single stems. We made the assumption that the original single stem RSV already provides an overall measure of the relevance of a document. Additional co-occurrence matches serve to improve this measure by a proportion. Thus, the new RSV for ranking becomes: RSV' = $(1+\alpha)$ * RSV, (with $0<=\alpha<1$). We limit matching patterns to exact phrase with highest $\alpha=w0$, and ordered or unordered matching within window size $w<5$ with $\alpha=w0/(w+0.5)$. For $w>=5$ up to a sentence, $\alpha=w0/5.5$. For each sentence only one tightest match is counted. For each document, repeat tightest matches are counted. We differentiate one from many by giving a repeat factor $r>1$ to the latter. Our RSV' then becomes $r*(1+w0/(w+0.5))$, $w=0..5$. Experimentation with various schemes show that this weight formula is simple and gives reasonable performance.

3. EXPERIMENTS & DISCUSSIONS
We studied 25 2-word queries from TREC-8 and 20 from TREC-7. An initial retrieval that does not employ adjacent 2-word phrases for representation, collection enrichment or other special techniques is used as basis (displayed as Col. 2 in Table 1). Standard TREC evaluation measures as labeled on the rows are used, emphasizing on precision at 10, 20, 30 documents retrieved since we do re-ranking. Cols. 3 & 4 show coordinate matching re-ranking of initial results with and without stemming. Cols. 5 & 6 are re-ranking results using our window weighting formula.

Contrary to [1], stemming or no-stemming does not lead to much difference. However, coordinate matching re-ranking is less effective than that based on our window formula. This is because these queries have only two words, and retrieved documents are often ordered by coordinate matching already during 1st stage retrieval. Window co-occurrence imposes tighter relevance evidence, and results improve when sentence context are also included. This window re-ranking method gives low recall precision (the mean of p@10, 20 and 30) improvements of about 4% for TREC-8 and 9-10% for TREC-7. This result does not involve 2-stage PRF, but can be useful in time-critical retrievals (such as in high volume web environment) where 2-stage retrieval may be too costly compared to re-ranking the top 200 documents of an initial retrieval.

We next look at how this improved initial retrieval may affect pseudo-relevance feedback. A fixed number of 40 terms were used for query expansion, but a varying number of top documents: 10, 20 and 30 were used. The original PRF results are displayed in Columns 7-9, and those involving co-occurrence re-ranking of 1st stage are shown in Columns 10-12. One may notice the original PRF (without re-ranking) for TREC-8 was not much better than the 1-stage retrieval until '30doc'. On the other hand, PRF retrieval after co-occurrence re-ranking provides more stable, better performance throughout. Re-ranking also considers the 'hardness of improvement' factor, i.e. several initial retrievals already have over 40% of their top 50 documents containing the exact query phrase, and these have their re-ranking proportion further reduced. Using the "30doc" column as example, co-occurrence re-ranking improves 2nd stage retrieval by about 8% for both TREC-8 and 7. Some of the tested queries are originally three-word, but reduced to two after high frequency thresholding. However, co-occurrence weighting made use of them in pairs. If the third words were not considered, the effectiveness would be reduced to 3% to 4%.

4. CONCLUSION

2-word queries provide the least complicated environment to study co-occurrence effects on retrieval. It is shown that it can improve retrieval of single stems by re-ranking using a simple inverse distance formula for window matching. Studying these queries may give hints on how to do better weighting for longer queries. This is a first step in our approach to use more individualized processes to optimize retrieval for different query types.

5. ACKNOWLEDGMENTS

This work was partially sponsored by the Space and Naval Warfare Systems Center San Diego, under Grant No. N66001-00-1-8912.

6. REFERENCES

[1] Crouch, C.J., Crouch, D.B., Chen, Q. & Holtz, S.J. Improving the retrieval effectiveness of very short queries. Info. Process. & Mngmt. 38(1) 2002, 1-36.

[2] Kwok, K.L. Improving English and Chinese ad-hoc retrieval: a Tipster Text Phrase 3 project report. Information Retrieval, 3(4) 2000, 313-338.

[3] Hawking, D. & Thistlewaite, P. Proximity operators – so near and yet so far. In: The Fourth Text Retrieval Conference (TREC-4). NIST SP 500-236, pp.131-144. GPO:D.C. 1996.

[4] Clarke, C.L.A, Cormack, G.V. & Burkowski, F.J. Shortest substring ranking (Multitext experiments for TREC-4). In: The Fourth Text Retrieval Conference (TREC-4). NIST SP 500-236, pp.295-304. GPO: Washington, D.C. 1966.

[5] Fagan, J.L. The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. JASIS 40, 1989, 115-132.

Table 1: Effect of Co-Occurrence Evidence on Performance of 2-Word Queries

Col. 1	2	3	4	5	6	7	8	9	10	11	12
Trec8	**Initial Stge-1**	**<------ coord stem**	**Rerank coord nostem**	**Stge-1 windw stem**	**--------> windw nostem**	**<------- 10doc**	**PRF original 20doc**	**-------> 30doc**	**<------ <after 10doc**	**PRF Stge-1 20doc**	**-------> Rerank> 30doc**
MAP	.2552	.2514	.2519	.2577	.2595	.2461	.2553	.2693	.2512	.2720	.2835
p@10	**.4640**	.4560	.4480	.4720	**.4640**	.4560	.4800	**.4840**	.4760	**.5480**	.5200
p@20	**.3940**	.3740	.3780	.4120	**.4060**	.3880	.3980	**.4420**	.4320	**.4580**	.4580
p@30	**.3453**	.3520	.3467	.3693	**.3760**	.3587	.3533	**.3760**	.3840	**.4013**	.3920
Trec7											
MAP	.2026	.2129	.2167	.2099	.2124	.2662	.2680	.2541	.2759	.2766	.2674
p@10	**.4400**	.4250	.4400	.4700	**.4700**	.4700	.5100	**.4900**	.5200	**.5150**	.4950
p@20	**.3650**	.3650	.3875	.4075	**.4075**	.3975	.4450	**.4250**	.4250	**.4625**	.4500
p@30	**.3200**	.3183	.3350	.3517	**.3600**	.3650	.3867	**.3667**	.3750	**.4083**	.3950