Statistical Phrases for Vector-Space Information Retrieval

Andrew Turpin Alistair Moffat

Department of Computer Science and Software Engineering
The University of Melbourne
Parkville, Victoria 3052, Australia
www.cs.mu.oz.au/~{aht,alistair}

1 Introduction

When employing a vector-space model to evaluate a query against a document collection several choices must be made. A fundamental design decision is the definition of the *terms* which form the dimensions of the space. Should the terms be single words, pairs of words, linguistic phrases, entire sentences, or some other combination of textual units? It seems intuitive that when calculating a measure of similarity between a natural language query text and natural language documents, some respect should be paid to word ordering. Complex terms such as phrases should, therefore, increase the precision of retrieval results. Recent work has, however, shown that this is not the case [8, 4]. In this abstract we describe experiments that further confirm that observation. Note that we are solely concerned with statistical phrases; that is, phrases derived using techniques other than NLP.

Exploration of phrases as terms in a vector-space based retrieval system has received detailed attention over at least 25 years. Salton et al. [6] show that including statistical phrases as terms in vector-space based retrieval increases precision averaged over 10 recall points by 17% to 39%. These experiments were updated by Fagan in 1989, who used larger document collections [2] (but, at about 10 MB, still small by today's standards). Fagan reports that average precision improvements range from −11% up to 20%. The downward trend in the impact of statistical phrases on average precision continued in 1997, with Mitra et al. [4] replicating Fagan's experiments on a 655 MB collection, and reporting a 1% precision improvement if phrases are used as terms. This surprising result is also supported in a separate study by Smeaton and Kelledy [8]. Our findings independently confirm these previous results, and add further evidence to the case against the use of phrases as precision-enhancing devices – a result that we still find somewhat surprising, since documents and queries are surely more than just bags of words.

2 A phrase

Mitra et al. [4] define a phrase to be any pair of non-function words that appear in at least 25 documents of the TREC-1 collection, with the two words sorted lexicographically. In an effort to replicate their experiments using our own system, a variant of the mg system that uses a word-level index, we defined a phrase to be any pair of words that occurs in the query, sorted lexicographically.

3 Methods

In order to closely study the effects of the various experimental design decisions on Mitra et al.'s results, we first attempted to replicate their study. We then removed several restrictions that they placed on phases, one by one, finding in each case that precision does not increase significantly.

The test collection employed was the concatenation of the Associated Press, Wall Street Journal, and Ziff-Davis components of the TREC-2 collection [3], a total of 211,359 documents. The query set was extracted from TREC topics 151 through 200, which is known to have at least 4,273 relevant documents in the AP2-WSJ2-ZIFF2 collection. Queries were constructed by removing stop words and mark-up tags from the concatenation of the "title", "description", and "narrative" components of each TREC topic.

We used a variation of the mg system that employs a word-level index [1] for retrieval. Storage of ordinal word positions allows the inverted list for a phrase to be constructed in memory during query processing. This gives us flexibility to study the effect of phrases without the need to re-index the collection when the definition of a phrase changes. All operations on terms were performed after stemming.

The similarity between a document and a query is determined by the Lnu.ltu formula, using SMART notation [7], also[1] known as BB-AGJ-BCA by Zobel and Moffat [9]. Seven pivot slopes between 0.05 and 0.70 were tested, 0.25 giving the best results, as it did for Singhal et al. The similarity value for a phrase is modified by a multiplicative phrase factor, which is "typically set at 0.5" [4]. Other phrase factors were tested, but we agree that 0.5 is superior.

4 Results

Table 1 shows the average precision results achieved by varying the definition of phrases, those reported by Mitra et al., and the precision achieved with the BD-ACI-BCA similarity measure, which Zobel and Moffat recommend for general purpose retrieval [9]. The precision values reported are determined by the trec_eval program [5] as the "Average precision (non-interpolated) over all rel docs". Statistical significance and p-values were determined by a Wilcoxon sign

[1] See www.cs.mu.oz.au/~alistair/irtools/smart2q.html for notes on how to convert SMART notation to a Q-expression.

SIGIR '99 8/99 Berkley, CA, USA
© 1999 ACM 1-58113-096-1/99/0007...$5.00

Similarity measure	Simple Terms	Simple terms & phrases	Percentage improvement	p-value (Wilcoxon rank sum)
Mitra et al.	0.3616	0.3758	3.9%	unknown
BD-ACI-BCA	0.3373	0.3579	5.76%	$p = 0.2060$
BB-AGJ-BCA	0.3479	0.3641	4.44%	$p = 0.031$
Unsorted pairs, occurring ≥ 25	0.3479	0.3654	4.77%	$p = 0.023$
Unsorted pairs, occurring ≥ 1	0.3479	0.3685	5.59%	$p = 0.005$
Unsorted, any length, occurring ≥ 1	0.3479	0.3660	4.93%	$p = 0.015$

Table 1: Precision averaged over all relevant documents, with 500 documents retrieved. Pivoting slope is 0.25, and the phrase factor is 0.5. The first row is derived from Mitra et al. [4].

rank sum test. A paired t-test was not used as Kolmogorov-Smirnov tests indicated that the distributions of the pairs of query runs were dissimilar ($p \geq 0.95$). [2]

The BB-AGJ-BCA row of Table 1 uses the same definition of phrases as Mitra et al., with the exception that phrases are only included as a term when they occur at least 25 times in the collection, rather than in the TREC-1 collection. This scheme falls short of the results achieved by Mitra et al., both with and without phrases. There are many possible reasons why this is the case. Firstly some of our phrases may not occur 25 times in the TREC-1 collection, but are included, and there may be some phrases that do not occur 25 times in AP2-WSJ2-ZIFF2 that do occur in TREC-1, hence are excluded. Secondly our stemming and stopping algorithms differ. In particular, we do not employ a stop list in the mg system, whereas the SMART system does. This affects the calculation of the average number of terms in a document, which in turn affects a term's document weight. The number of unique terms in each document is also altered, which impacts document normalisation. Adjusting these values (using the SMART system's stop list where necessary) improves precision marginally, but did not reverse the poor performance of phrases.

There are doubtless other subtle differences between SMART and the mg-based system employed in our experiments. What is encouraging, however, is that BB-AGJ-BCA outperforms BD-ACI-BCA (shown in the first row), the method recommended for general purpose retrieval by Zobel and Moffat [9], which indicates that our implementation is a good approximation of Lnu.ltu under the SMART system.

Including phrases leads to a 4.44% improvement in precision for our system, a greater percentage gain than attained by Mitra et al., but to a lesser final precision. This adds credence to the claim by Mitra et al. that the reason the impact of phrases as terms has dropped over the last 25 years is that the precision achieved using simple terms has increased dramatically. With the mg-based mechanism we start with a lower baseline precision figure (using BB-AGJ-BCA) of 0.3497, and there is greater scope for phrases to have an impact.

The final three rows of Table 1 show precision when the definition of a phrase is altered from the BB-AGJ-BCA method. Row four uses unsorted pairs that occur at least 25 times, row five uses unsorted pairs that occur at least once, and row six uses phrases of any length, unsorted, occurring at least once. All three increase precision slightly, but not significantly ($p = 0.05$). Hence, the only benefit of our approach is that we construct phrases "on the fly" rather than choosing phrases at index construction time, and are not limited by the initial resource costs associated with indexing the phrases of a collection.

5 Conclusions

We agree with Mitra et al. that using non-NLP generated phrases as terms in vector-space retrieval are not the precision enhancing devices that they intuitively should be. We further observed that using phrases helped at low recall levels, but did not help in the top (say) 20 documents retrieved; a trend also observed by Mitra et al.

We have experimented with other relevance measures, other collections, long and short queries, and varying definitions of phrases. In each case similar results arise. As the reader has probably surmised, we were unable to identify any use of phrases that resulted in substantial improvements in precision. There is, however, still scope for treating phrases in a separate vector-space from words, as suggested by Smeaton and Kelledy [8].

References

[1] O. de Kretser and A. Moffat. Effective document presentation with a locality-based similarity heuristic. In *Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, August 1999. To appear.

[2] J. L. Fagan. The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science*, 40(2):115–132, 1989.

[3] D. K. Harman. Overview of the first text retrieval conference. In D. K. Harman, editor, *Proc. TREC Text Retrieval Conference*, pages 1–20, Washington, November 1992. National Institute of Standards Special Publication 500-207.

[4] M. Mitra, C. Buckley, A. Singhal, and C. Cardie. An analysis of statistical and syntactic phrases. In *Proceedings of RIAO97, Computer-Assisted Information Searching on the Internet*, pages 200–214, Montreal, Canada, June 1997.

[5] C. Buckley. trec_eval program, March 1999. Available from ftp.cs.cornell.edu/pub/smart.

[6] G. Salton, C. S. Yang, and C. T. Yu. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1):33–44, 1975.

[7] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson, editors, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–29, Zurich, Switzerland, August 1996. ACM Press, NY.

[8] A. F. Smeaton and F. Kelledy. User-chosen phrases in interactive query formulation for information retrieval. In *Proceedings of the 20th BCS-IRSG Colloquium*, Springer-Verlag Electronic Workshops in Computing, Grenoble, France, April 1998.

[9] J. Zobel and A. Moffat. Exploring the similarity space. *SIGIR Forum*, 32(1):18–34, 1998.

[2] Software for performing these tests on trec_eval output is available at www.cs.mu.oz.au/~alistair/irtools/.