

# Evaluating the Potential of Explicit Phrases for Retrieval Quality

Andreas Broschart<sup>1,2</sup>, Klaus Berberich<sup>2</sup>, and Ralf Schenkel<sup>1,2</sup>

<sup>1</sup> Saarland University, Saarbrücken, Germany

<sup>2</sup> Max-Planck-Institut für Informatik, Saarbrücken, Germany

**Abstract.** This paper evaluates the potential impact of explicit phrases on retrieval quality through a case study with the TREC Terabyte benchmark. It compares the performance of user- and system-identified phrases with a standard score and a proximity-aware score, and shows that an optimal choice of phrases, including term permutations, can significantly improve query performance.

## 1 Introduction

Phrases, i.e., query terms that should occur consecutively in a result document, are a widely used means to improve result quality in text retrieval [2,3,4,5,7], and a number of methods has been proposed to automatically identify useful phrases, for example [5,11]. However, there are studies indicating that phrases are not universally useful for improving results, but that the right choice of phrases is important. For example, Metzler et al.[6] reported that phrase detection did not work for their experiments in the TREC Terabyte track, and Mitra et al. [8] reported similar findings for experiments on news corpora.

This paper experimentally analyses the potential of phrase queries for improving result quality through a case study on the TREC Terabyte benchmark. We study the performance improvement through user-identified and dictionary-based phrases over a term-only baseline and determine the best improvement that any phrase-based method can achieve, possibly including term permutations.

## 2 Experimental Setup

We did a large-scale study on the effectiveness of phrases for text retrieval with the TREC GOV2 collection, a crawl of approximately 25 million documents from the .gov domain on the Web, and the 150 topics from the TREC Terabyte AdHoc tracks 2004-2006 (topics 701–850). All documents were parsed with stopword removal and stemming enabled. We compared different retrieval methods:

- A standard BM25F scoring model [9] as an established baseline for content-based retrieval, with both conjunctive (i.e., all terms must occur in a document) and disjunctive query evaluation.

- Phrases as additional post-filters on the results of the conjunctive BM25F, i.e., results that did not contain at least one instance of the stemmed phrase were removed. As the TREC topics don't contain explicit phrases, we considered the following ways to find phrases in the queries:
  - We performed a small user study where five users were independently asked to highlight any phrases in the titles of the TREC queries.
  - As an example for a dictionary-based method for phrase detection, we matched the titles with the titles of Wikipedia articles (after stemming both), following an approach similar to the Wikipedia-based phrase recognition in [11].
  - To evaluate the full potential of phrases, we exhaustively evaluated the effectiveness of all possible phrases for each topic and chose the best-performing phrase(s) for each topic.
  - To evaluate the influence of term order, we additionally considered all possible phrases for all permutations of terms and chose the best-performing phrases, potentially after permutation of terms, for each topic.
- A state-of-the-art proximity score [1] as an extension of BM25F, including the modifications from [10]; this score outperformed other proximity-aware methods on TREC Terabyte according to [10].

We additionally report the best reported results from the corresponding TREC Terabyte tracks, limited to title-only runs.

### 3 Results

Our small user study showed that users frequently disagree on phrases in a query: On average, two users highlighted the same phrase only in 47% of the queries, with individual agreements between 38% and 64%. For each topic with more than one term, at least one user identified a phrase; for 43 topics, each user identified a phrase (but possibly different phrases). The same user rarely highlighted more than one phrase in a topic. Overall, our users identified 227 different phrases in the 150 topics.

Our experimental evaluation of query effectiveness focuses on early precision. We aim at validating if the earlier result by [8] (on news documents) that phrases do not significantly improve early precision is still valid when considering the Web. Table 1 shows precision@10 for using the phrases identified by the different users (as strict post-filters on the conjunctive BM25F run). Surprisingly, it seems to be very difficult for users to actually identify useful phrases, there hardly is any improvement. In that sense, the findings from [8] seem to be still valid today.

In the light of these results, our second experiment aims at exploring if phrase queries have any potential at all for improving query effectiveness, i.e., how much can result quality be improved when the 'optimal' phrases are identified. Tables 2 and 3 show the precision@10 for our experiment with the different settings introduced in the previous section, separately for each TREC year.

It is evident from the tables that an optimal choice of phrases can significantly improve over the BM25F baseline, with peak improvements between 12% and

**Table 1.** Precision@10 for user-identified phrases

|         | BM25F<br>(conjunctive) | user 1 | user 2 | user 3 | user 4 | user 5 |
|---------|------------------------|--------|--------|--------|--------|--------|
| 701-750 | 0.536                  | 0.512  | 0.534  | 0.504  | 0.546  | 0.536  |
| 751-800 | 0.634                  | 0.576  | 0.484  | 0.548  | 0.592  | 0.602  |
| 801-850 | 0.528                  | 0.518  | 0.500  | 0.514  | 0.546  | 0.526  |
| average | 0.566                  | 0.535  | 0.506  | 0.522  | 0.561  | 0.554  |

**Table 2.** Precision@10 for different configurations and query loads, part1

|         | BM25F<br>(disjunctive) | BM25F<br>(conjunctive) | best user<br>phrases | Wikipedia<br>phrases |
|---------|------------------------|------------------------|----------------------|----------------------|
| 701-750 | 0.548                  | 0.536                  | 0.546                | 0.566                |
| 751-800 | 0.630                  | 0.634                  | 0.592                | 0.564                |
| 801-850 | 0.538                  | 0.528                  | 0.546                | 0.526                |
| average | 0.572                  | 0.566                  | 0.561                | 0.552                |

14% when term order remains unchanged, and even 17% to 21% when term permutations are considered<sup>1</sup>. Topics where phrases were most useful include “pol pot” (843), “pet therapy” (793) and “bagpipe band” (794) (which were usually identified by users as well). On the other hand, frequently annotated phrases such as “doomsday cults” (745) and “domestic adoption laws” (754) cause a drastic drop in performance. Interesting examples for improvements when permuting terms are “hybrid alternative fuel cars” (777) where the best phrase is actually “hybrid fuel” (with a P@10 of 0.8, compared to 0.5 for the best in-order phrase and 0.2 for term-only evaluation), and “reintroduction of gray wolves” (797) with p@10 of 1.0 with the phrase “wolves reintroduction”, compared to 0.6 otherwise.

**Table 3.** Precision@10 for different configurations and query loads, part2

|         | proximity<br>score | best<br>phrases | best phrases<br>incl. permutations | best title-only<br>TREC run |
|---------|--------------------|-----------------|------------------------------------|-----------------------------|
| 701-750 | 0.574              | 0.616           | 0.668                              | 0.588                       |
| 751-800 | 0.660              | 0.704           | 0.740                              | 0.658                       |
| 801-850 | 0.578              | 0.606           | 0.654                              | 0.654                       |
| average | 0.604              | 0.642           | 0.687                              | 0.633                       |

The best possible results are way above the best reported results for 2004 and 2005 and get close to the best result from 2006 (which was achieved, among other things, by the use of blind feedback)<sup>2</sup>. Wikipedia-based phrase recognition, a simple automated approach to phrase recognition, does not lead to significant improvements. Interestingly, the proximity-aware score yields significant

<sup>1</sup> Both significant according to a t-test with a p-value  $\leq 0.01$ .

<sup>2</sup> No significance tests possible as we don’t have per-topic results for these runs.

improvements over the baseline<sup>3</sup>; as it automatically considers “soft phrases”, there is no need to explicitly identify phrases here.

## 4 Discussion and Lessons Learned

The experimental analysis done in this paper yields the following results:

- We validated the common intuition that phrase queries can boost performance of existing retrieval models. However, choosing good phrases for this purpose is nontrivial and often too difficult for users, as the result of our user study shows.
- Existing methods for automatically identifying phrases can help to improve query performance, but they have their limits (like the methods based on Wikipedia titles evaluated here). While we expect that more complex methods such as the advanced algorithm introduced in [11] will get close to the upper bound, they need to include term permutations to exploit the full potential of phrases. The common intuition that term order in queries bears semantics does not seem to match reality in all cases.
- Proximity-aware scoring models where the user does not have to explicitly identify phrases can significantly improve performance over a non-proximity-aware scoring model.

## References

1. Büttcher, S., Clarke, C.L.A., Lushman, B.: Term proximity scoring for ad-hoc retrieval on very large text collections. In: SIGIR, pp. 621–622 (2006)
2. Clarke, C.L.A., Cormack, G.V., Tudhope, E.A.: Relevance ranking for one to three term queries. In: RIAO, pp. 388–401 (1997)
3. Croft, W.B., Turtle, H.R., Lewis, D.D.: The use of phrases and structured queries in information retrieval. In: SIGIR, pp. 32–45 (1991)
4. Fagan, J.L.: Automatic phrase indexing for document retrieval: An examination of syntactic and non-syntactic methods. In: SIGIR, pp. 91–101 (1987)
5. Liu, S., Liu, F., Yu, C.T., Meng, W.: An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In: SIGIR, pp. 266–272 (2004)
6. Metzler, D., Strohman, T., Croft, W.B.: Indri trec notebook 2006: Lessons learned from three terabyte tracks. In: TREC (2006)
7. Mishne, G., de Rijke, M.: Boosting web retrieval through query operations. In: Losada, D.E., Fernández-Luna, J.M. (eds.) ECIR 2005. LNCS, vol. 3408, pp. 502–516. Springer, Heidelberg (2005)
8. Mitra, M., Buckley, C., Singhal, A., Cardie, C.: An analysis of statistical and syntactic phrases. In: RIAO, pp. 200–217 (1997)
9. Robertson, S.E., Zaragoza, H., Taylor, M.J.: Simple BM25 extension to multiple weighted fields. In: CIKM, pp. 42–49 (2004)
10. Schenkel, R., Broschart, A., Hwang, S.-W., Theobald, M., Weikum, G.: Efficient text proximity search. In: Ziviani, N., Baeza-Yates, R. (eds.) SPIRE 2007. LNCS, vol. 4726, pp. 287–299. Springer, Heidelberg (2007)
11. Zhang, W., et al.: Recognition and classification of noun phrases in queries for effective retrieval. In: CIKM, pp. 711–720 (2007)

---

<sup>3</sup> p-value  $\leq 0.1$  for TREC 2005 and  $\leq 0.01$  for the other two.