Boosting Web Retrieval through Query Operations

Gilad Mishne and Maarten de Rijke

Informatics Institute, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
E-mail: {gilad,mdr}@science.uva.nl

Abstract. We explore the use of phrase and proximity terms in the context of web retrieval, which is different from traditional ad-hoc retrieval both in document structure and in query characteristics. We show that for this type of task, the usage of both phrase and proximity terms is highly beneficial for early precision as well as for overall retrieval effectiveness. We also analyze why phrase and proximity terms are far more effective for web retrieval than for ad-hoc retrieval.

1 Introduction

An important aspect in which web retrieval differs from ad-hoc retrieval concerns the users needs. User studies and anecdotal evidence suggest that web users wish to spend as little time as possible going through the results, and are mostly interested in a small number of relevant documents in the topmost ranks. Most users look only at the first page of results (usually, containing 10 results) [20, 32], and this trend is strengthening over time [31]. Moreover, web search users usually have short search sessions, indicating that once a user followed a link to a document which she finds relevant, she will in most cases not return to the result list and examine further hits [7].

Accordingly, recent large-scale web search evaluations such as the web track at TREC [10, 11] have widened the traditional focus on evaluation measures such as Mean Average Precision (MAP) and Precision/Recall graphs to also include early precision based measures such as Precision@10, Precision@20 and Success@10; in some cases, even higher precision is evaluated, e.g., Mean Reciprocal Rank (MRR, mostly for tasks with a single relevant document) and Precision@5, Success@5 and Success@1. The latter measures are motivated by the fact that, due to physical limitations, the first 10 results are not always displayed in a single "screen page," requiring the user to scroll down the list.

The web continues to be an inspiring domain for retrieval research. For instance, the layout information embedded within HTML documents gave rise to numerous refinements and extensions of retrieval models that attempt to take non-content features of documents into account [8]. Our focus in this paper is not on web retrieval models but on web *queries*. How can we boost web retrieval effectiveness, measured using any of the measures just mentioned, by means of automatic operations on queries?

An important difference between web retrieval queries and typical queries in other retrieval tasks is the average query length. Web search user studies such as those mentioned earlier report on average lengths of 1.5 to 2.6 terms; similar numbers have recently been cited by top web search engines [25] and also emerge out of web query logs we are currently gathering. In contrast, closed-domain searches have significantly higher average lengths, e.g., 4.9 terms for the TREC 2004 Genomics track [18]. Given these observations on query length, it is obviously important to make the most out of what little information web queries give us. We examine the effect of automatic query rewrites, specifically phrasal and proximity-based retrieval, on the performance of web retrieval. A phrase match between a document and a query is usually an accurate indication that the document deals with the aspect of the query described by the phrase. Intuitively, the ability to detect overlap between a document and a query aspect is particularly important if queries are short and may have very few aspects.

We are especially interested in the effectiveness of "light-weight" query operations for web retrieval. Thus, we do not consider phrases as indexing units, but submit queries that exploit phrases or proximity terms against an index consisting of single terms only. Also, our phrases are not syntactic or even statistical in nature; we simply treat every word n-gram from the query as a phrase. For us, proximity based retrieval is a natural extension of phrasal retrieval where the restriction on the nearness of the terms is somewhat more relaxed.

Now, usage of proximity and phrases has been studied extensively for ad-hoc retrieval. Reports on their contribution are mixed, and it is generally accepted now that with a good basic ranking formula, the effectiveness of phrases is negligible or even negative [24], while recent evaluations of the use of automatically generated proximity terms suggest that term proximity may improve retrieval effectiveness especially at the top documents retrieved [28]. Our main research questions are:

- Given a good basic ranking scheme for web retrieval, how much additional benefit do phrases and proximity terms bring in retrieval effectiveness?
- To what extent are improvements gained by phrases and proximity terms dependent on the structured nature of web documents?
- Do phrases and proximity terms impact Mean Average Precision scores differently than high precision measures?
- Do phrases and proximity terms have a different impact on retrieval effectiveness for extremely short queries (2 or 3 terms) than for longer queries?

One of our main findings is that because of the structured nature of web documents, phrases and proximity terms can increase effectiveness for web retrieval. When using short (or very short) queries to retrieve HTML documents, significant improvements can be obtained if phrases and proximity terms are used, not only in terms of the high-precision measures mentioned above but, interestingly, also in terms of traditional measures such as Mean Average Precision.

The rest of the paper is organized as follows. In Section 2 we survey work on phrasal retrieval, discuss current web retrieval efforts, and describe state of the art techniques used for the latter task. In Section 3 we describe the phrase and

proximity based methods we experimented with for boosting web retrieval effectiveness; we motivate them, and give examples. Next, in Section 4, we describe our experimental framework, largely based on the TREC web track retrieval evaluations. We follow with an account of our results, comparing them to the performance of other techniques for web retrieval. In Section 5 we provide a deeper analysis for some topics, aiming to understand where our methods are especially beneficial or detrimental to web retrieval effectiveness. Finally, our conclusions and ongoing work come in Section 6.

2 Background and Related Work

Web Retrieval. In recent years, web retrieval tasks were divided into two categories: *Named Page Finding* and *Topic Distillation*. Named page finding targets scenarios where a user searches for a specific page (which is known to exist, such as a personal home page); this task is often evaluated with MRR or Success@N for low values of N, since the user is known to be interested in only one result, and prefers it to be as high on the ranked list as possible. Topic distillation, on the other hand, involves finding key resources for a particular subject. Distillation is normally evaluated with traditional MAP and precision@N scores [10, 11].

We focus on retrieval for topic distillation. Why? First, current performance on the named page task is very high, making it almost a solved problem. In the 2003 edition of the TREC web track, top performing systems achieved 90% Success@10 and 0.7 MRR scores for this task [11], meaning that in most of the cases the single relevant document is returned at rank 1. Furthermore, the median scores over all participating systems are 80% for Success@10 and well over 0.5 for MRR. In contrast, the topic distillation task has lots of room for improvements: at TREC 2003, the best performing system scored less than 0.13 on Precision@10 and less than 0.16 on MAP.

Secondly, the good results on named page retrieval are partly due to the heavy usage of factors not directly related to the ranking formula (e.g., indegree information); this makes the task highly sensitive to these external factors, thus making it more complex to study the effects of changes in the ranking algorithm or query processing on retrieval performance.

Finally, we focus on topic distillation because we want to determine the impact of the use of phrases and proximity terms both in terms of the traditional MAP scores and in terms of (very) high precision measures such as MRR, Precision@1/Precision@5, and Success@1/Success@5. Topic distillation is unique as a task where both types of evaluation scores make sense.

Phrases and Proximity Terms. Intuitively, proximity and phrase operators are factors which improve retrieval effectiveness; indeed, lots of research was directed in this direction. The relative merits of statistical and syntactic phrases were extensively investigated by Fagan [14], and again by Hull *et al.* [19]. Until the late 1990s, usage of phrases and proximity operators—as well as a careful usage of boolean operators—did show varying degrees of improvements of retrieval results [17, 12, 22], but rarely anything substantial.

As retrieval models became more advanced, the usage of various query operators was questioned. Mitra *et al.* [24] investigate the effectiveness of using phrases for plain text retrieval (on a standard newswire text collection); they employ both linguistic and statistical methods for phrase extraction. Their conclusion is that when using a good, modern ranking algorithm, phrases have no effect on high precision retrieval (and sometimes negative effect from topic drift); for low precision, there is some marginal improvement from the usage of phrases. Similar conclusions have been reached for non-English IR, also on plain text [23].

Work on retrieval using a proximity framework is more scarce. Hawking and Thistlewaite explore the use of proximity scoring within the PADRE system [16]. Clarke and Cormack [9] show promising results, especially for manually-refined queries; it is unclear how this approach is combined with $tf \cdot idf$ based models, which constitute the majority of today's retrieval approaches (including Okapi and Language Modeling, which usually derive the estimations used in them from these factors). Rasolofo and Savoy [28] combine term-proximity scoring heuristics with the Okapi probabilistic model, obtaining 3%–8% improvements for Precision@5/10/20, with hardly observable impact on the MAP scores.

There has been relatively little systematic work on the effectiveness of phrases and proximity terms in the setting of web retrieval. At the TREC 2003 web track, however, several participants reported improvements based on proximity information, spans, and phrases [11]; two of the five top performing systems in the named page finding task used proximity in some way [30, 33]. However, we were unable to find systematic evaluations of the use of proximity terms in queries compared to the same ranking formula with no use of proximities.

Our work on query operations differs from earlier work because of our exclusive focus on web retrieval, exploiting the structure of web documents as well as the special content of some document fields (such as URLs and anchors), and because of our focus on "light-weight" phrases that are computationally cheap and robust against grammatical and spelling errors often found in web queries.

3 Query Refinement for Web Retrieval

In this section we describe the operations we use for query refinement and motivate their selection as an approach for improving web retrieval effectiveness.

3.1 Phrases and Proximity Terms

Previous research on the use of phrases for query refinement discusses statistical, syntactical, and lexical phrase detection [3, 14, 24, 27]. All approaches show mixed results on ad-hoc retrieval, with the maximal gain to precision being 5%–7%. We follow a different, shallow way of phrase detection: an "everything-is-a-phrase" approach. In our view, phrase terms need not necessarily be actual phrases, either in the syntactical or statistical sense; they can simply be words which appear consecutively in relevant documents, with high likelihood. For example, for topic WT04-58 from TREC 2004, "automobile emissions vehicle

pollution," it seems that many subsets of consecutive words from the query are relevant as phrase terms, regardless of the statistical or syntactical evidence for their "phrasehood." Such subsets are "automobile emissions" and "vehicle pollution" but also "emissions vehicle" (which matches, after stopping and stemming, "emissions from a vehicle" or "emitted by vehicles"). While this also creates non-phrases, linguistically or statistically, the frequency of such word n-grams in the collection is virtually zero [6], preventing performance degradation. So, in our experiments, we choose to consider every word n-gram (of any length, inclusive single words and all words) which is part of the query, as a phrase. This naive approach carries with it some practical benefits: robustness, low computational overhead, no noise created by additional mechanisms and algorithms, etc.

For proximity operators, we employ a similar approach. We consider all word n-grams from the query as a proximity term; we then experiment with two query rewriting methods to exploit proximities: *fixed distance* and *variable distance*. Using the fixed distance method, every n-gram is a proximity term with a fixed distance, which depends on the length of the n-gram and an externally provided parameter. For example, if the parameter is $k = 2$, the n-gram is "emissions vehicle", and the method for combining the parameter and the length is multiplying them, the distance we have for this proximity is 4. We experimented with estimation methods for deriving the proximity distance from the external parameter and the n-gram length, e.g., linear combinations, products, squared combinations, and so on; we found no major differences in average performance (for both early precision and overall performance measures), provided that the values of k are tuned for the specific combination with the n-gram length. Hence, we use a simple sum of the external parameter and the size of the n-gram; the value of k was empirically set to 11. This type of combination allows longer proximity terms a larger distance, loosening the restrictions on longer terms which tend to be ungrammatical (e.g., "automobile emissions vehicle pollution").

With the fixed distance method, assuming the calculated distance is n, all occurrences of the term words in windows of n and less are scored equally. To reward terms according to the actual distance between the proximity terms, within the variable distance rewriting method we rewrote a proximity term into a series of proximity terms, each having a lower distance restriction. Terms which are found in smaller windows than n will match more than one term, effectively increasing the ranking of the document. Practically, this is done using the same method used to generate the fixed distance proximity terms, but with decreasing values of k. For example, the term "automobile emissions vehicle pollution" will be translated into 11 separate query terms, ranging from a fixed distance term with $k = 11$ down to the same term with $k = 1$.

In all our experiments, the result list was reranked using link indegree and URL length as reported in [1].[1]

[1] We note that similarly to the results obtained there, the reranking substantially improved all measures, up to 60% improvement in early precision scores. The improvements seemed consistent for all models—with or without usage of query operators—and we consider them orthogonal to the results of the various query reformulations.

3.2 Query Operators in the Vector Space Model

In our experiments, we focus on the vector space model, for which all advanced query operators are well researched; virtually any IR textbook (e.g., [4, 29]) contain a discussion of operators such as phrases, proximity, and wildcards. Rather than tuning up the retrieval formula, tweaking it to match the specific task that is addressed, we use a fixed, basic ranking formula. For this formula, we define the ranking of both simple terms and more complex ones (e.g., phrase terms). We then experiment with a range of transformation methods for deriving terms out of the original query; the definition given for ranking each term type is used to derive the final ranking formula.

Given a collection D, the basic similarity score between a document d and a query q containing terms t_i in our experiments is a common vector space variation:

$$sim(q,d) = \sum_{t \in q} \frac{tf_{t,q} \cdot idf_t}{norm_q} \cdot \frac{tf_{t,d} \cdot idf_t}{norm_d} \cdot coord_{q,d} \cdot weight_t,$$

where

$$tf_{t,X} = \sqrt{\text{freq}(t,X)} \qquad idf_t = 1 + \log \frac{|D|}{\text{freq}(t,D)}$$
$$norm_d = \sqrt{|d|} \qquad coord_{q,d} = \frac{|q \cap d|}{|q|}$$
$$norm_q = \sqrt{\sum_{t \in q} tf_{t,q} \cdot idf_t{}^2}$$

Terms can be either a single word, a phrase, or a proximity term. For single term words, the *tf* and *idf* calculation is straightforward. For a multiple-word term t (phrase or proximity), composed of the single word terms $t_0, t_1, \ldots, t_n$, the actual frequency counts in the collection of a phrase are not normally used, mainly for efficiency reasons. There are various ways to estimate these figures; previous experiments have shown little difference in performance between methods [24]. We experimented with the following estimation methods, testing early precision measures as well as MAP:

`Sum`:	$idf = \sum_{i=0}^{n} idf_i$
`Minimum`:	$idf = \min_i idf_i$
`Maximum`:	$idf = \max_i idf_i$
`Arithmetic Mean`:	$idf = \sum_{i=0}^{n} idf_i / n$
`Geometric Mean`:	$idf = \prod_{i=0}^{n} idf_i{}^{\frac{1}{n}}$

The results, evaluated on the test set described in Section 4.1, are presented in Table 1; best scores (for a given evaluation measure) are in boldface. As may be seen from the results in Table 1, for some measures there are differences between estimation methods, but they are not dramatic. As an aside, phrase terms seem to display more variability than proximity terms. We choose the `Minimum` estimation, which seems to provide good performance both for early precision and for overall precision scores. The `Minimum` estimation also seems more intuitive, since phrase occurrences should be more restrictive than the occurrences of the words within them.

Table 1: Comparison of *idf* estimation methods.

	Phrases			Proximity		
Method	P@10	S@10	MAP	P@10	S@10	MAP
`Sum`	0.1576	0.7440	0.1438	**0.1888**	0.7600	**0.1569**
`Minimum`	**0.1712**	0.7680	0.1433	0.1832	**0.7840**	0.1502
`Maximum`	0.1688	0.7600	**0.1457**	0.1832	0.7840	0.1502
`Arithmetic Mean`	**0.1712**	**0.7760**	0.1450	0.1824	0.7760	0.1485
`Geometric Mean`	**0.1712**	**0.7760**	0.1433	0.1824	0.7760	0.1482

As for the *tf* figures for multiple-word terms, they remain the same as single-word ones, i.e. real frequencies of the multiple-word term in the document or the query. The frequency is calculated according to the multiple-word restrictions, e.g., if the term is a proximity term with two single word terms in a span of 10 words, an "occurrence" of it will be counted every time the two words appear in a window of 10 words or less. For example, in the document "dog cat mouse dog dog cat", the number of occurrences of the phrase "dog cat" is 2, and the proximity term "dog cat" with distance 3 has 4 occurrences.

3.3 Multiple Representations of Documents

When addressing web retrieval, most of the target documents are HTML documents containing markup, rather than simple plain text. This markup has been extensively used in the web retrieval setting, for example by top performers in the TREC web retrieval tasks, to form a more sophisticated document representation than a bag-of-words (see e.g., [2, 26]).

We make use of the markup by dividing each document into multiple "fields" which are indexed separately, providing separate frequency estimates for each field. The fields we identify in an HTML document are TITLE, DESCRIPTION, KEYWORDS, BODY, URL and ANCHOR TEXT. We experimented extensively with the use of different combinations of fields; our best results consistently appear when using only the TITLE, BODY and ANCHOR TEXT field. We attribute the lack of contribution of the DESCRIPTION and KEYWORDS fields to the relative sparseness of their usage: only 16% of the documents in our corpus (described in Section 4.1) contain the META keyword "description" and only 18% contain the "keywords" keyword.

Additionally, we experimented with methods for assigning term weights to the phrase terms according to the length of the n-gram, various external parameters and hard-coded assumptions (e.g., "TITLE is more important than URL"). For the majority of the methods, the effect on performance was not substantial. We did establish consistent if small improvements when using term weights derived from the real frequencies of the term (as a phrase) in a certain field in the collection, and report on this in Section 4.

4 Evaluation

In this section we describe our experiments and their outcomes.

Table 2: Distribution of topic lengths.								
Topic Length	1	2	3	4	5	6	**Mean**	2.38
Topic Count	10	64	25	11	4	2		

4.1 Experimental Setting

We follow the experimental setup of the web tracks at TREC 2003 [11] and TREC 2004 [13]. The corpus used for the experiments is the `.GOV` corpus, which is a crawl of a subset of the `.gov` domain performed in 2002. The corpus contains 18.1Gb of data in 1.25M documents, the vast majority of which are HTML documents, and it preserves the link information between the documents. Our test set consists of the two topic distillation topic sets released with TREC 2003 and 2004, containing 50 and 75 queries respectively, for a total of 125 queries, with topic lengths as detailed in Table 2. We use the assessments provided by the organizers of the web tracks.

4.2 Experiments and Results

First, we provide a brief description of the different query formulation methods we experimented with.

- `baseline`: All words from the topic are single-word terms.
- `phrases`: All word n-grams from the topic are used as phrase terms, as described in Section 3.1.
- `phrases-b` Same as `phrases`, but every phrase term is given a term weight proportional to the real term frequency of the term phrase (as a phrase) in different fields.
- `proximity` All word n-grams from the query are used as proximity terms, with a fixed distance length.
- `prox-v` All word n-grams from the query are used as proximity terms, with a variable distance length.

The scores of the different experiments for early precision measures and additional measures are presented in Table 3 and Table 4, respectively.

On the TREC 2003 distillation topics, the baseline achieves scores which would position it among the top 10 experiments (out of 93 experiments) for all measures; for Precision@10, the baseline equals the best reported score. Our non-baseline runs score better than any reported experiment.

4.3 Discussion

Mitra *et al.* [24] report that the use of phrases yields little or no improvement, provided that the basic ranking formula is a good one. When using a single field representation of the document, i.e., all text—title, body, propagated anchor text and so on—is indexed in the same field, we reach similar conclusions. Interestingly, however, for the multiple field representation of documents, we clearly

Table 3: Comparison of methods, early precision measures.

Method	P@10	P@5	S@10	S@5	MRR
Single field representation					
`baseline`	0.1456	0.1840	0.7040	0.5440	0.4193
`phrases`	0.1456 (0%)	0.1888 (+2%)	0.7200 (+2%)	0.5520 (+1%)	0.4273 (+2%)
`proximity`	0.1528 (+5%)	0.1968 (+7%)	0.7280 (+3%)	0.5900 (+8%)	0.4126 (-2%)
`prox-v`	0.1488 (+2%)	0.2064 (+12%)	0.7200 (+2%)	0.5940 (+9%)	0.4283 (+2%)
Multiple field representation					
`baseline`	0.1720	0.2224	0.7520	0.6400	0.4811
`phrases`	0.1712 (-1%)	0.2288 (+3%)	0.7680 (+2%)	0.6240 (-2%)	0.5215 (+8%)
`phrases-b`	0.1912 (+11%)	0.2416 (+9%)	0.7600 (+1%)	0.6560 (+2%)	0.4992 (+4%)
`proximity`	0.1888 (+10%)	0.2512 (+13%)	0.7920 (+5%)	0.6560 (+2%)	0.5142 (+7%)
`prox-v`	0.1904 (+11%)	0.2496 (+12%)	0.7840 (+4%)	0.6560 (+2%)	0.5156 (+7%)

see an improvement on all measures when using phrases and proximities, up to 23% on some measures. Observe, moreover, that these improvements cannot be attributed to a low baseline: as pointed out before, the baseline achieves state of the art performance on the 2003 topics, and well above median performance on the 2004 ones. Additionally, our non-baseline runs score better, on some early precision measures, than unrelated state of the art models we use for the task [21].

Concluding that our baseline is sufficiently high, we take a closer look at the results. A number of observations can be made. First, clearly, the less restrictive the additional operators are, the larger the improvement is to performance: fixed proximity terms outperform the more rigid phrase term, but are themselves generally not as good as the flexible proximity terms. Second, the use of plain phrases, without the additional term weights, yields unstable results—improving some measures but degrading others. Finally, the Success@10 measure is the most difficult to improve, possibly since it is high to start with.

Combinations. There is strong evidence suggesting that combinations of different retrieval techniques results in significant improvement of results (see, e.g., [5]).

Table 4: Comparison of methods, additional measures.

Method	R-Prec	MAP
Single field representation		
`baseline`	0.1157	0.1041
`phrases`	0.1235 (+6%)	0.1088 (+4%)
`proximity`	0.1282 (+10%)	0.1094 (+5%)
`prox-v`	0.1267 (+9%)	0.1101 (+5%)
Multiple field representation		
`baseline`	0.1578	0.1271
`phrases`	0.1687 (+7%)	0.1433 (+13%)
`phrases-b`	0.1607 (+2%)	0.1443 (+13%)
`proximity`	0.1791 (+13%)	0.1569 (+23%)
`prox-v`	0.1822 (+15%)	0.1559 (+22%)

Table 5: Performance comparison for most frequent query lengths.					
Topic	**Topic**	**Phrases**		**Proximity**	
Length	**Count**	P@10	MAP	P@10	MAP
2	64	0.2286 (21%)	0.1392 (22%)	0.2254 (19%)	0.1451 (27%)
3	25	0.1640 (17%)	0.1295 (23%)	0.1520 (8%)	0.1658 (57%)
4	11	0.1273 (-18%)	0.1195 (-25%)	0.1455 (-6%)	0.1499 (-5%)

Since we used different query modifications, we had reason to believe that combinations of them are worthwhile; we therefore experimented with various ways to combine between our experiments. Additionally, we combined the results of our methods with a completely different set of experiments, based on the language modeling approach to IR and achieving in itself very good results for web IR at TREC 2003 and 2004 [1]. We observed that combinations yield consistent improvements of an additional 3%–5% percent both to early precision and to average precision measures. For space reasons, we do not report on the experiments here, and will give a more detailed account in [21].

5 Topic Analysis

In this section we provide a more detailed analysis of the impact of phrase and proximity operators on the retrieval effectiveness of individual queries. To save space, we restrict our discussions and comparisons to experiments using the multiple field representation; moreover, the results we obtained on the single field representation are similar to ones already reported by others.

5.1 Topic Length and Effectiveness

The most visible factor determining the effectiveness of the phrasal and proximity methods in our experiments is, not surprisingly, the query length. The mean length of the topics in our test set is 2.38, in line with the average query length for web retrieval mentioned earlier (Section 1). In Table 5 we examine the Precision@10 and MAP scores separately for different topics lengths, and their change from the baseline. We do not include topics of length 1 (for which there is no change in the ranking formula), and topics of length 5 and above, which constitute only 4% of the topics and are not statistically significant.

We can observe a strong correlation between the length of the queries and the effectiveness gain: the gain is significantly larger for topics of relative short (2–3) length. This is largely due to the fact that many of the shorter (2–3 term) queries were formed of a single linguistic phrase, whereas longer queries are commonly just a collection of words. For longer lengths than those displayed in the table (such as 5 or 6 words) we observed an even larger drop in performance.[2]

[2] The results in Table 5 suggest that phrases and proximity should not be used for topics of length 4 or more. We experimented with the "best" setting for each group of topics (where topics are grouped by length). As topics of length 4 or more account

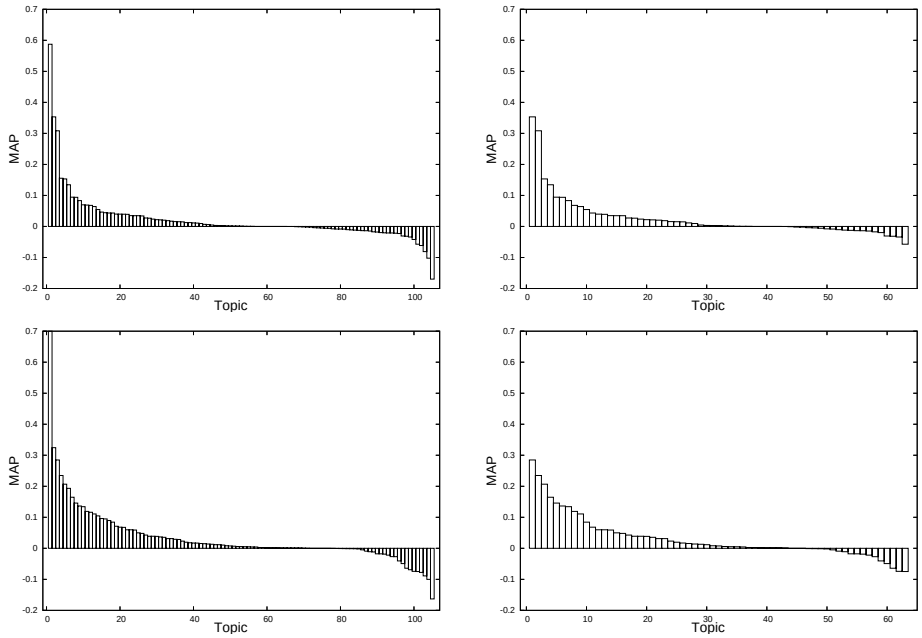

Fig. 1. Per-topic gain in Mean Average Precision compared to the baseline. (Top row): Using phrases, with all 106 topics longer than 1 word (left) and all 64 topics of length 2 (right). (Bottom row): Using proximity terms, with all 106 topics longer than 1 word (left) and all 64 topics of length 2 (right).

A further breakdown of individual gain per topic is given in Figure 1 (effect on MAP for phrases and proximity) and Figure 2 (effect on Precision@10 for phrases and proximity). The histograms show similar behavior of both phrases and proximity terms. Improvements are generally far greater and far more frequent than degradations. When looking at topics of all lengths, 10% to 20% have a significant improvement, another 30%–40% some improvement, and for about 30% of the topics the usage of the query operators results in reasonably limited reduced effectiveness. Results for the 2-word topics are analogous, with larger percentages of topics achieving effectiveness gains.

5.2 Examples

In Table 6 we take a closer look at the scores of a number of specific topics from the test set; in addition to Precision@10 scores, we list Average Precision (AP) scores per topic. A discussion regarding the causes for the differences in effectiveness for each topic follows.

for less than 14% of the topics, no dramatic differences could be observed with the results in Table 4.

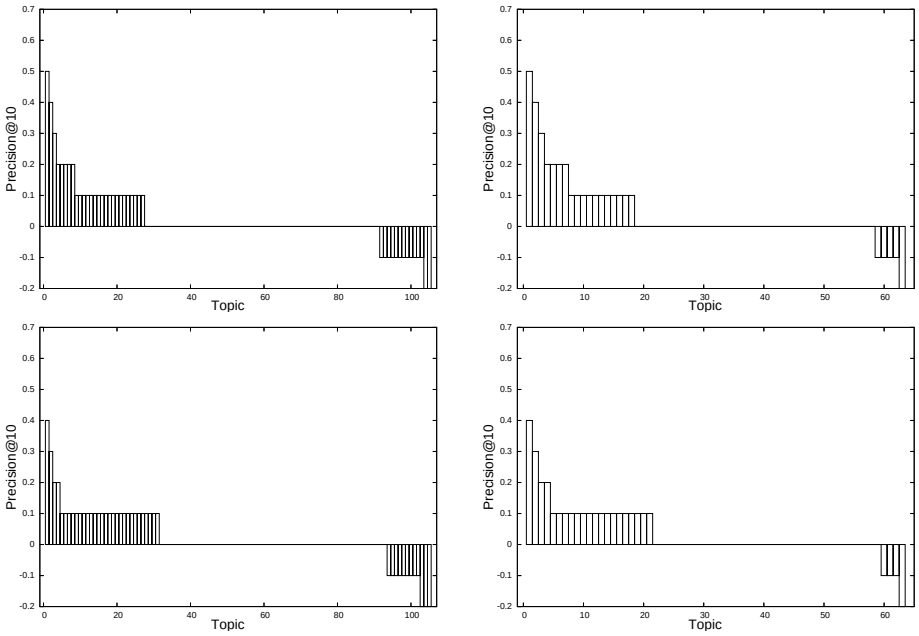

Fig. 2. Per-topic gain in Precision@10 compared to the baseline. (Top row): using phrases, with all 106 topics longer than 1 word (left) and all 64 topics of length 2 (right). (Bottom row): using proximity terms, with all 106 topics longer than 1 word (left) and all 64 topics of length 2 (right).

Table 6: Individual topic examples.

	Baseline		**Phrases**		**Proximity**	
Topic	P@10	AP	P@10	AP	P@10	AP
`skin cancer`	0.2	0.1208	0.4	0.3350	0.4	0.3275
`homeland security`	0.1	0.0721	0.3	0.2065	0.3	0.2064
`diet nutrition weight management`	0.2	0.1107	0.0	0.0297	0.1	0.0840
`deafness in children`	0.2	0.0903	0.1	0.0917	0.1	0.1190

The first two topics, `skin cancer` and `homeland security`, are somewhat classical examples of the effectiveness of using proximity between terms in the ranking. In the baseline model, the score is heavily dominated by the term `cancer` and `security`, which appear in short fields such as title and anchor text. In this case, both the usage of proximity and of phrases yields significant improvements.

With the third topic, `diet nutrition weight management`, we encounter the opposite effect, with the baseline scores being better than the other methods. Here, many of the relevant documents had different phrases than those appearing in the query; such phrases are “weight loss”, “weight control”, and so on. In this

case, proximity terms have a better performance since the constraints they pose on the document are more liberal, compensating for the lexical gap by pushing up documents in which the query terms are close.

Performance on the final topic, `deafness in children`, is similar to the previous one, i.e., the lack of improvement by phrases is attributed to a "phrase lexical gap." Phrases which are common in relevant documents, e.g., "hearing loss," "hard of hearing," "assistive listening systems," etc, do not appear in the query. However, since the query is shorter, the effect is less dramatic.

6 Conclusions and Ongoing Work

Earlier studies on the use of phrases and proximity terms show little improvement, particularly when the base ranking is good. Our experiments show that for web retrieval this is not the case. For this task, the combination of short, focused queries with documents that contain short, focused information (e.g., HTML titles) leads to significant improvements when using those query operators, even when applied in a naive fashion. The performance gains can be observed both for early precision mesures and for mean average precision.

The phrasal and proximity methods seem to help more the shorter the queries are; the queries that gain the most have, on average, the same length as the average length of web search engine queries. For longer queries, these methods cause topic drift, and need to be applied more carefully; we leave this issue for future work.

Returning to our main research questions, as formulated in Section 1, we have found that even on top of a good basic ranking scheme for web retrieval, phrases and proximity terms may bring improvements in retrieval effectiveness. While we observed improvements both when documents are represented as a single field, and as aggregates of multiple fields, the latter setting gave more substantial improvements. Somewhat suprisingly, we found that phrases and proximity terms improve scores for traditional mean average precision as well as for high precision measures, although the former tended to be more substantial. Another important finding was that phrases and proximity terms have a strong positive impact on web retrieval effectiveness for extremely short queries (2 or 3 terms), while they have less, or even negative, effects on longer queries.

We are currently exploring approaches to the usage of phrases and proximity terms in the language modeling framework for web retrieval. We expect that the theoretic foundations of language modeling will provide a better understanding of how and where usage of these operators improves effectiveness. Additionally, we will apply our current results to additional corpora where, similarly to web documents, multiple representations of documents exist: such corpora are XML documents [15] and biomedical document collections [18].

Acknowledgments

We wish to thank Jaap Kamps for many discussions and inspiration, and our referees for useful comments.

Both authors were supported by the Netherlands Organization for Scientific Research (NWO) under project number 220-80-001. Maarten de Rijke was also supported by NWO under project numbers 365-20-005, 612.069.006, 612.000.106, 612.000.207, 612.066.302, 264-70-050, and 017.001.190.

References

[1] D. Ahn, V. Jijkoun, J. Kamps, G. Mishne, K. Müller, M. de Rijke, and S. Schlobach. The University of Amsterdam at TREC 2004. In *TREC 2004 Conference Notebook*, Gaithersburg, Maryland USA, 2004.

[2] E. Amitay, D. Carmel, A. Darlow, M. Herscovici, R. Kraft, R. Lempel, A. Soffer, and J. Zien. Juru at TREC 2003 - Topic Distillation using Query-Sensitive Tuning and Cohesiveness Filtering. In *Proceedings of the 12th Text REtrieval Conference*, 2003.

[3] A.T. Arampatzis, Th.P. van der Weide, C.H.A. Koster, and P. van Bommel. An Evaluation of Linguistically-motivated Indexing Schemes. In *Proceedings of the 22nd BCS-IRSG Colloquium on IR Research*, 2000.

[4] R.A. Baeza-Yates and B.A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[5] B.T. Bartell, G.W. Cottrell, and R.K. Belew. Automatic Combination of Multiple Ranked Retrieval Systems. In *Research and Development in Information Retrieval*, pages 173–181, 1994.

[6] E. Brill, S. Dumais, and M. Banko. An analysis of the AskMSR question-answering system. In *Proceedings 39th Annual ACL*, 2002.

[7] F. Cacheda and A. Vina. Understanding how people use search engines: a statistical analysis for e-business. In *Proceedings of the e-Business and e-Work Conference and Exhibition*, pages 319–325, Venice, Italy, Oct 2001.

[8] S. Chakrabarti. *Mining the Web: Analysis of Hypertext and Semi Structured Data*. Morgan Kaufmann, 2002.

[9] C.L.A. Clarke and G.V. Cormack. Shortest-substring retrieval and ranking. *ACM Transactions on Information Systems (TOIS)*, 18(1):44–78, 2000.

[10] N. Craswell and D. Hawking. Overview of the TREC-2002 web track. In *Proceedings of TREC-2002*, Gaithersburg, Maryland USA, November 2002.

[11] N. Craswell, D. Hawking, R. Wilkinson, and M. Wu. Overview of the TREC-2003 web track. In *Proceedings of TREC 2003*, Gaithersburg, Maryland USA, November 2003.

[12] W.B. Croft, H.R. Turtle, and D.D. Lewis. The use of phrases and structured queries in information retrieval. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 32–45. ACM Press, 1991.

[13] N. Craswell et al. Overview of the TREC-2004 web track. In *Proceedings 13th Text REtrieval Conference*, Gaithersburg, Maryland USA, To appear.

[14] J.L. Fagan. Experiments in automatic phrase indexing for document retrieval: A comparison of syntactic and non-syntactic methods. Technical report, Cornell University, 1987.

[15] N. Fuhr, M. Lalmas, and S. Malik, editors. *INEX 2003 Workshop Proceedings*, 2004.

[16] D. Hawking and P. Thistlewaite. Proximity operators—So near and yet so far. In *Proceedings TREC-4*, pages 131–143, 1996.

[17] D. Hawking and P. Thistlewaite. Relevance weighting using distance between term occurrences. Technical Report TR-CS-96-08, Department of Computer Science, Australian National University, 1996.

[18] W. Hersh and R.T. Bhupatiraju. TREC GENOMICS Track Overview. In *Proceedings TREC 2003*, pages 14–23, 2004.

[19] D.A. Hull, G. Grefenstette, B.M. Schultze, E. Gaussier, H. Schutze, and J. O. Pedersen. Xerox TREC-5 Site Report: Routing, Filtering, NLP, and Spanish Tracks. In *Proceedings TREC-5*, pages 167–180, 1997.

[20] B.J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, 36(2):207–227, 2000.

[21] J. Kamps, G. Mishne, and M. de Rijke. The University of Amsterdam at TREC 2004. In *Proceedings of the 13th Text REtrieval Conference*, to appear.

[22] E.M. Keen. Term position ranking: some new test results. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 66–76. ACM Press, 1992.

[23] W. Kraaij and R. Pohlmann. Comparing the effect of syntactic vs. statistical phrase index strategies for Dutch. In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, pages 605–617, 1998.

[24] M. Mitra, C. Buckley, A. Singhal, and C. Cardie. An analysis of statistical and syntactic phrases. In *Proceedings of RIAO-97*, 1997.

[25] V. Mittal, S. Baluja, and M. Sahami. Google tutorial on web information retrieval. In *RIAO-2004*, 2004.

[26] P. Ogilvie and J. Callan. Combining document representations for known-item search. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM Press, 2003.

[27] J. Pickens and W.B. Croft. An exploratory analysis of phrases in text retrieval. In *Proceedings of RIAO-2000*, 2000.

[28] Y. Rasolofo and J. Savoy. Term proximity scoring for keyword-based retrieval systems. In *Proceedings 25th European Conference on IR Research (ECIR 2003)*, pages 207–218, 2003.

[29] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986.

[30] J. Savoy, Y. Rasolofo, and L. Perret. Report on the TREC-2003 experiment: Genomic and web searches. In *Proceedings TREC 2003*, pages 739–750, 2004.

[31] A. Spink, B.J. Jansen, D. Wolfram, and T. Saracevic. From e-sex to e-commerce: Web search changes. *Computer*, 35(3):107–109, 2002.

[32] A. Spink, D. Wolfram, B.J. Jansen, and T. Saracevic. Searching the web: the public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234, 2001.

[33] J. Wen, R. Song, D. Cai, K. Zhu, S. Yu, S. Ye, and W.-Y. Ma. Microsoft Research Asia at the Web Track of TREC 2003. In *Proceedings TREC 2003*, pages 408–417, 2004.