Boosting Titles does not Generally Improve Retrieval Effectiveness

Jimmy
Queensland University of Technology, Australia
& University of Surabaya, Indonesia
jimmy@hdr.qut.edu.au

Guido Zuccon
Queensland University of Technology, Australia
g.zuccon@qut.edu.au

Bevan Koopman
Australian e-Health Research Centre, CSIRO, Australia
bevan.koopman@csiro.au

ABSTRACT

The fields that compose structured documents such as web pages have been exploited to improve the effectiveness of information retrieval systems. Field-based retrieval methods assign different levels of importance (weights) to different fields, e.g., by boosting the score of a document when query terms are found in a specific field. An important question is how to decide which field should be boosted? It has been speculated that the title field should receive a higher weight. In this paper, we investigate whether boosting the title field of structured documents actually does improve retrieval effectiveness. Our results show that, on average, boosting titles does not improve retrieval effectiveness for field-based retrieval; this is both for ad-hoc web search and exploratory-based web search tasks. However, we do find that the boosting of titles does generally improve retrieval effectiveness for navigational queries and a small subset of ad-hoc queries. This result advocates for adaptive methods that selectively adjust boosting of specific fields based on the query.

1. INTRODUCTION

Web pages are structured text documents [13, 6]: information contained in a web page is organized into standardized fields such as title, headers, keywords, body, etc. Each field is used by the author of the web page for different purposes. Titles are used to briefly convey and emphasize the content of the page. The body collects the content aimed at the reader. While, fields like "keywords" and "description" are not meant for the reader but for web search engines and other computerized programs.

The structure of web pages (and other structured documents, e.g., XML files) has been long exploited in information retrieval by devising retrieval models that weight or combine evidence (e.g., keyword matches) from different fields in different ways [18, 17, 11, 9, 22]. For example, BM25F [17] extends the popular BM25 retrieval model by weighting matches in different fields according to "boosting factors" assigned individually to each field. Field-based retrieval models generally improve the effectiveness of retrieval systems [17].

In this context, a key question is what weights should be assigned to the different fields? That is, which field(s) should be boosted, and by how much? This is an important question because, aside from large commercial search engines like Google and Bing that rely on sophisticated learning-to-rank algorithms and a wide range of search interaction signals, there exists a large array of search systems that still rely on standard best match models such as BM25(F), e.g. search services within an organization. A clear answer to this question, however, does not exist. It has been speculated that the weight assigned to the title field should be boosted above that assigned to the body field (main content of a document) [7, 18]. This is because a query matching the title of a document may provide stronger evidence of relevance than an equivalent match on the body. Put in other words, "a title [is expected] to be much denser in topic-specific terms than an average body sentence" [17]. This speculation is supported by work by Joho et al. [7], who report that top ranked documents in web search tend to have query terms in the title. Similarly, matches on the title field are often associated with higher weights in machine learning approaches for web search such as learning-to-rank [5]. Boosting titles is also the commonly assumed to-do for tuning the open source search engine ElasticSearch/Lucene, e.g., see https://www.elastic.co/guide/en/elasticsearch/guide/current/query-time-boosting.html and http://www.lucenetutorial.com/lucene-query-syntax.html.

In this paper, we question this speculation regarding boosting titles. While queries issued to satisfy navigational intents may indeed be best answered by retrieval systems that boost title matches over matches in other fields, queries associated to exploratory needs are often unlikely to exhibit useful matches with documents' titles. Instead, in such cases title matches and body matches should be considered equivalent or, in some cases, body matches should be boosted. To further study this, we experiment with a field-based retrieval system by varying the boost weights assigned to selected web page fields. We consider two types of search tasks: queries related to general, ad-hoc search (on the web and on newswire) and queries related to more exploratory information needs, as represented by the consumer health search task (i.e., average people seeking health advice online [25, 15]). Specifically, we aim to address the following research questions:

ADCS '16, December 05 - 07, 2016, Caulfield, VIC, Australia

ACM ISBN 978-1-4503-4865-2/16/12...$15.00

DOI: http://dx.doi.org/10.1145/3015022.3015028

RQ1: Do top ranked documents contain the query terms in the title field?

This question aims to extend the work of Joho et al. [7]. They, in fact, considered the same question and reported that top ranked documents (by both commercial web search engines and baseline IR models) tend to have a high proportion of query terms in their title, called *query-in-title* (QIT). Here, we want to verify whether this is the case both in general web and newswire search, and in more exploratory tasks such as consumer health search. Our hypothesis is that in exploratory tasks like consumer health search the QIT ratios are lower than for general web or newswire search. Intuitively, this is because exploratory queries are often circumlocutory and tend to contain descriptive terms rather than more concise and technical terms that are more likely to appear in titles (for example the technical name of a condition, e.g., Hypertension, rather than the more circumlocutory alternative, e.g., high blood pressure) [25, 20].

RQ2: Does boosting the title field over the body improve retrieval effectiveness? Does this hold for all types of search, ad-hoc web vs. exploratory?

This question challenges the common assumption in previous work that the title field should be boosted above the body in field-based retrieval models. Our hypothesis is that in exploratory tasks such as consumer health search, boosting the title would not lead improvements in retrieval effectiveness when compared to treating body and title as equivalent, or even boosting the body field.

2. RELATED WORK

Next we provide a brief account of prior work related to our two research questions. Specifically, we first examine methods that exploit the structure of a web page to improve retrieval effectiveness. We then examine the use of document title as an important feature within the retrieval process.

2.1 Exploiting the Structure of a Web Page

The structure of a web page can be successfully exploited to improve the effectiveness of information retrieval systems [18, 13, 19, 22].

A retrieval model that does this is BM25F, which extends the standard BM25 retrieval model by defining boosting factors associated to matches in different fields of a document [16]. In this paper, we focus on BM25F because this model is commonly employed in standard search engines (e.g., Lucene and Elastic Search [16]) and because its scores are often used as a feature to inform web learning-to-rank algorithms [5]. Zaragoza et al. [24] formally defined BM25F as:

$$BM25F(d) := \sum_{t \in q \cap d} \frac{\bar{x}_{d,t}}{K_1 + \bar{x}_{d,t}} w_t^{(1)} \quad (1)$$

where:

$$\bar{x}_{d,t} = \sum_f W_f \cdot \bar{x}_{d,f,t} \quad (2)$$

$$\bar{x}_{d,f,t} := \frac{x_{d,f,t}}{(1 + B_f(\frac{l_{d,f}}{l_f} - 1))} \quad (3)$$

In Equation 3, f indicates the document field type (e.g., body, title, anchor text), and $x_{d,f,t}$ represents the term frequency of term t in the field type f for document d. Further, $l_{d,f}$ is the length of field f in document d and l_f is the average length of that field type.

In this $BM25F$ function, we need to define one normalization parameter (B_f in Equation 3) and one weight parameter (W_f in Equation 2) for each field. Only one saturation parameter K_1 is required and is applied to all fields. In our experiments, we shall investigate the impact of W_f on retrieval effectiveness, while leaving B_f constant for all fields and, as for K_1, set according to common values used in the literature.

An alternative way of exploiting field information in the BM25 model has been suggested by Robertson [18] and consists in repeating the content of each field based on the weight assigned to that field and then combine the repeated fields into a single unstructured document. In this approach, assigning a weight of 2 to the title field and 3 to the body field is equivalent to create a surrogate document with the title repeated twice and the body repeated three times.

In the language modelling framework, fields can also be modelled to inform retrieval [14]. This is done by using a two-step generation process: the first step measures the likelihood of the query generating the selected field; the second step measures the likelihood of the field to generate the document. As with BM25F, the language modelling framework also allows for fields to be weighted individually according to their importance. We defer the empirical study of field-based language modelling for IR in the context of our research questions to future work. However, we note that trends observed for BM25 based experiments are often also found in language modelling experiments.

Molinari et al. [13] have suggested a method based on term statistics and distribution of terms in fields to determine appropriate weights for the individual fields. The study of whether this method would allow one to predict the best field boosting (as we shall empirically observe in our experiments) is left for future work. However, as we do show in Section 4, optimal field weighting does appear to be dependent on the tasks and the query, rather than collection statistics.

In this work we focus on the body and the title weights. Previous work has shown the importance of link anchor information in improving retrieval effectiveness, especially for navigational queries [4]. We defer the investigation of these and other fields to future work.

2.2 The Role of Document Title

Joho et al. [7] speculate that the title field has the greatest influence when ranking documents. This is supported by the fact that they found top-ranked documents contained the query terms in the title and that users principally view page titles (and snippets) when presented with search results. This latter consideration is supported by the empirical studies of Clarke et al. [2], who show that people tend to select results that contain query terms in the title.

Empirical studies have shown that boosting the title leads to improved retrieval effectiveness. For example, Robertson et al. [18] reported improvements in precision at 10 when boosting title. Xu et al. [23] also used titles to improve effectiveness by adding title matches as a boosting factor in the retrieval model. In learning-to-rank, title information, including its BM25F weight, is often used as a feature to in-

form the learning method, and this is generally found to be a strong indicator of relevance [5]. Although there has been an attempt to understand the contribution title provides to retrieval effectiveness, little work has evaluated the contribution other fields may make if boosted, or the comparative contribution title has over boosting other fields, e.g., body.

3. EXPERIMENT SETUP

We conduct a number of retrieval experiments based on a set of representative IR test collections for general web search tasks and exploratory search tasks.

The general web test collections are TREC 2013 and 2014 Web Track collections [8, 3] (WEB2013-2014). These collections consist of 50 queries each (100 queries in total) and we use the Clueweb12-B13 corpus, which contains a crawl of more than 52 million web pages.[1] We also include the TREC 2005 HARD Track test collection (HARD2005) [1], which contains 50 queries and uses the AQUAINT corpus, a dataset of over 1 million newswire documents. This collection is used because it was the original collection used to show the effectiveness of BM25F, and the collection used by Joho et al. on their study of query-in-title [7] we shall further develop here.

As test collections representing exploratory search tasks, we use the CLEF2015 [15] and CLEF2016 [26] datasets, which focus on consumer health web search. CLEF2015 contains 66 queries and these were evaluated against the Khresmoi corpus which contains more than 1 million health related web pages. CLEF2016 contains 300 queries and these are evaluated against the Clueweb12-B13 corpus.

The corpora are parsed using python's lxml.html library[2] and four fields are extracted for indexing: title, meta, headers, and body. The title field contains the text between the tags `<title>` and `</title>`. The meta field contains the content of meta tags named "keywords" and "description". The headers field contains the content of `<h1>` to `<h6>` tags; lastly, the body field contains the text between the `<body>` and `</body>` tags. The corpora are further pre-processed by discarding documents without body, removing escape sequences, HTML tags and special characters (e.g., `&amp;`), and replacing non-alphanumeric characters with space (e.g., "self-esteem" is transformed to "self esteem"). Since the AQUAINT corpus does not have meta and headers fields, we only fetched its headline (i.e., title) and text (i.e., body).

After parsing, cleaning, and splitting the corpus content into fields, we index the corpora using ElasticSearch version 2.3.4[3], a popular production-rated open source search engine. We use BM25F as matching function with b=0.75 and $K_1 = 1.2$ that are the default parameter values in ElasticSearch and in many IR experiments. For indexing, we lowercase all text, stemmed each term using Porter's English stemmer and removed stop-words using the stop-words list supplied with Terrier [12].

We did not optimize the BM25F parameters (i.e., B_f and K_1), neither overall nor by field as suggested by Robertson et al. [18]; this tuning may improve retrieval effectiveness, but we leave the study of how this affects our findings for future work. Note that in many real world cases, search engine providers, for example at small corporate level, may not have enough data and assessments to reliably fine tune the BM25F parameters to specific values for each fields: thus the values used as default by the search engines may be their best bet.

For the retrieval experiments, we use a two-tailed t-test for identifying statistical significant differences between the effectiveness of different settings of field weightings. The statistical significance is reported when results are presented in table form.

[1]We added to the Clueweb12-B13 corpus 40 documents from the larger Clueweb12 corpus that are not in B13 but that have been assessed as navigationally relevant in the TREC 2013 and 2014 Web Track (these were all the documents with a navigational relevant assessment). This was done to study the role of title boosting for navigational queries (Section 4.3).

[2]http://lxml.de/

[3]http://www.elastic.co/

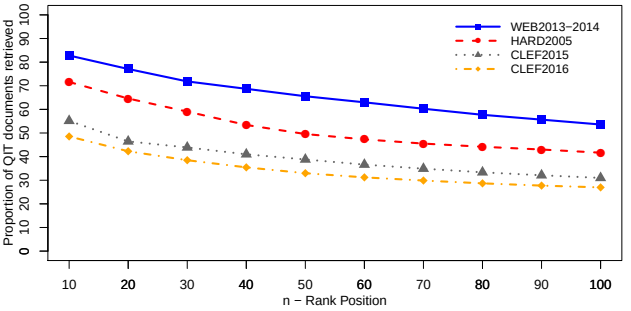


Figure 1: Proportion of QIT documents in top n results ($10 \leq n \leq 100$).

Collection	Relevant QIT	Not Relevant QIT
WEB2013-2014	6.79%	6.08%
HARD2005	24.30%	13.96%
CLEF2015	23.78%	20.87%
CLEF2016	14.13%	5.70%

Table 1: Proportion of QIT in relevant and not relevant documents.

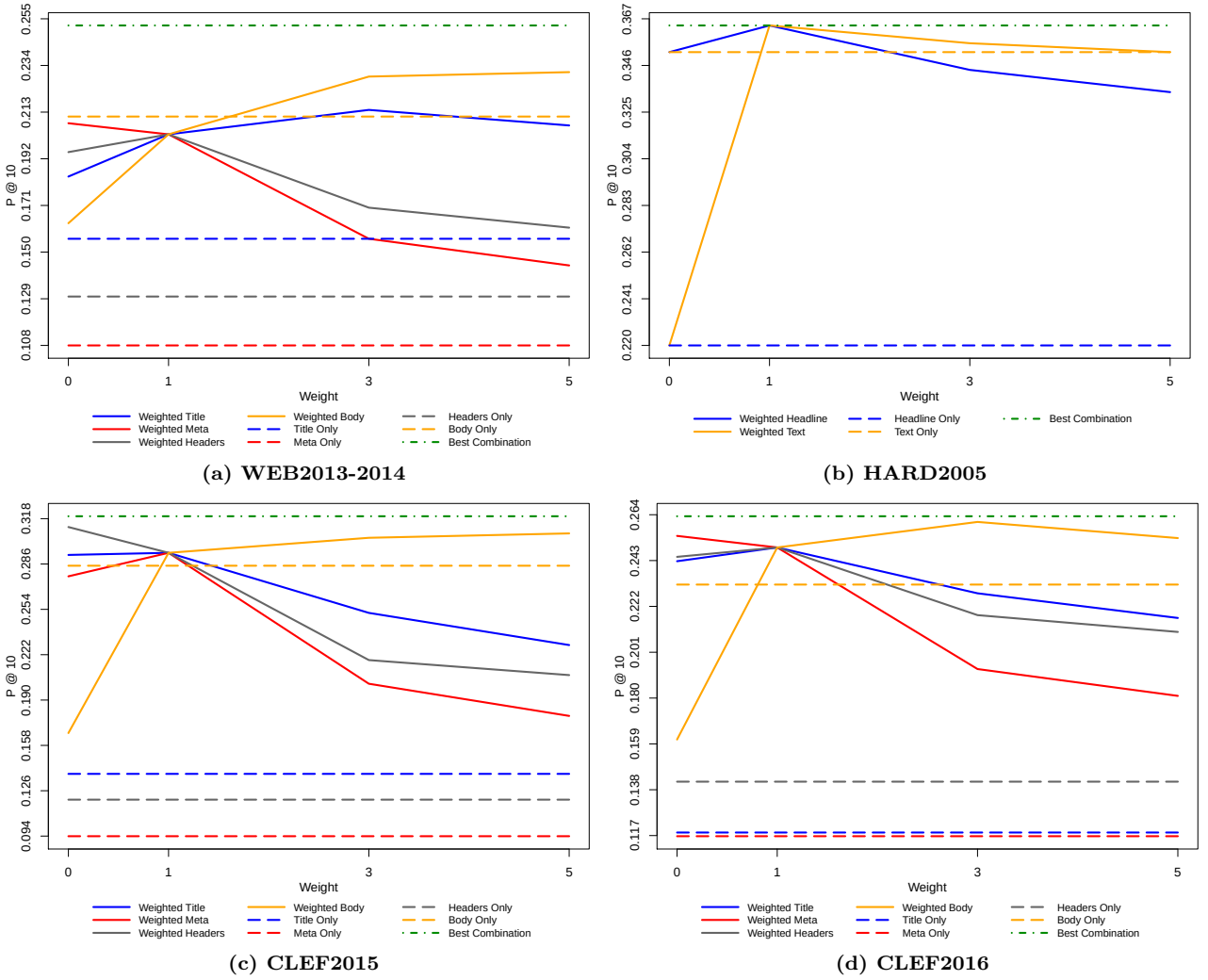
4. EXPERIMENTS AND RESULTS

4.1 Does the Title of Top Search Results Match the User Query?

We address the first research question by extending the work of Joho et al. [7]. Their research showed that most of the top 20 documents retrieved by search engines contain a query term in the title. To further study this, we measure the amount of *Query-in-Title* (QIT) [7] using topics and relevance judgments from the collections; in these experiments fields are weighted equally.

Figure 1 shows the proportion of documents in the top n results ($10 \leq n \leq 100$) with at least one query term in the title. More exploratory search tasks (CLEF2015 and CLEF2016) are characterized by significantly lower QIT values than general web search tasks (WEB2013-2014) or ad-hoc newswire search (HARD2005).

We also measure the proportion of QIT in documents explicitly judged relevant and not relevant by the assessors that created the collections. Table 1 shows that the proportion of QIT in documents judged relevant is higher than the proportion of QIT in not relevant documents.



(a) WEB2013-2014 **(b) HARD2005**

(c) CLEF2015 **(d) CLEF2016**

Figure 2: Retrieval effectiveness (p@10) of the field-based retrieval approaches when different weights are assigned to fields. *Field Only* approach refer to when the weight of the named field is set to 1 and all other weights are zero. *Weighted Field* approach refer to then the weight of the named field is varied from 0 to 5, and all other weights are set to 1.

4.2 Impact of Field Boosting

Next, we study the impact of field boosting on retrieval experiments (RQ2). When computing IR evaluation measures, we regard unassesed documents as irrelevant, except for BPREF that only considers judged documents.

Figure 2a shows the highest values of p@10[4] obtained on WEB2013-2014 for different field weighting combinations. The figure shows the results for when only the content of a specific field is used for retrieval, e.g., *TitleOnly* corresponds to set all the field weights to zero but title, which is set to 1. The figure also shows a study of varying the weights of a specific field, when all other fields are set to 1, i.e. when a specific field is boosted over the others. For example for *WeightedTitle* we vary the weight of the title from 0 to 5, maintaining the weights of the other fields to 1. Finally, *BestCombination* refers to the highest value achievable by letting each field weight vary across the explored range. For WEB2013-2014 the best combination is obtained when title and body are boosted by a weight of 3 and 5, respectively, and any other field is ignored.

The results in Figure 2a suggest that for web search, contrary to what commonly assumed, boosting the title field does not improve retrieval effectiveness and the highest retrieval effectiveness in terms of p@10 is obtained when instead the body field is boosted (*WeightedBody*), or when both fields are boosted but body is boosted higher (*BestCombination*). Similar results are obtained for HARD2005 (Figure 2b), where generally title and body fields need to be equally weighted to obtain the highest P@10 values.

The results for CLEF2015 and CLEF2016, i.e. the more exploratory web search tasks, are shown in Figures 2c and 2d.

[4] p@10 was used as primary measure in the CLEF collections; HARD2005 official evaluation was also based on a precision oriented measure (R-precision), while WEB2013-2014 also based its evaluation on top 10 documents, but with graded measures. Results with graded measures are reported later in the paper, along with a measure of the overall ranking quality (MAP)

	WEB2013-2014			HARD2005		
	nDCG@10	MAP	BPREF	nDCG@10	MAP	BPREF
Title Only	0.1183^{cde}	0.0196^{cde}	0.0612^{bcde}	0.1971^{bcde}	0.0509^{bcde}	0.1378^{bcde}
Body Only	0.1440^{e}	0.0250^{ce}	0.0831^{ace}	0.2869^{a}	0.1638^{ade}	0.2217^{ae}
Title = Body	0.1654^{a}	$\mathbf{0.0329}^{abd}$	$\mathbf{0.0879}^{abd}$	$\mathbf{0.3013}^{a}$	0.1646^{a}	$\mathbf{0.2289}^{ad}$
Title > Body	0.1533^{a}	0.0277^{ac}	0.0806^{ace}	0.2914^{a}	0.1227^{abce}	0.2085^{ac}
Body > Title	$\mathbf{0.1711}^{ab}$	0.0316^{ab}	0.0863^{abd}	0.2898^{a}	$\mathbf{0.1687}^{abd}$	0.2270^{ab}

	CLEF2015			CLEF2016		
	nDCG@10	MAP	BPREF	nDCG@10	MAP	BPREF
Title Only	0.1108^{bcde}	0.0364^{bcde}	0.1195^{bcde}	0.1040^{bcde}	0.0268^{bcde}	0.0973^{bcde}
Body Only	0.2729^{ad}	0.1758^{ad}	$\mathbf{0.2660}^{a}$	0.1968^{ace}	0.0753^{ace}	0.1495^{acde}
Title = Body	0.2793^{ad}	0.1749^{ad}	0.2615^{ad}	$\mathbf{0.2169}^{abd}$	$\mathbf{0.0854}^{abde}$	0.1613^{abe}
Title > Body	0.2030^{abce}	0.1085^{abce}	0.2215^{ace}	0.1954^{ac}	0.0707^{ace}	$\mathbf{0.1645}^{abe}$
Body > Title	$\mathbf{0.2800}^{ad}$	$\mathbf{0.1804}^{ad}$	0.2654^{ae}	0.2096^{ab}	0.0803^{abcd}	0.1550^{abcd}

Table 2: MAP, nDCG@10 and BPREF values of selected field weighting approaches: in all collections, boosting the body field, or weighting the body field equally to the title field, is more effective than boosting the title field. An exception is found for CLEF2016 using BPREF suggesting unassessed documents may influence the findings in this collection. Superscripts ^{a,b,c,d,e} represent that there are statistical significant differences ($p < 0.05$) between the result at the methods Title Only, Body Only, Title = Body, Title > Body, and Body > Title, respectively.

The results for these collections show similar trends as those for general web and ad-hoc search, and in particular that body weights have more influence on retrieval results than title weights, and that the best results are obtained when both these fields are boosted over the other fields, e.g., boosting both body and titles by a weight of 3, while the rest is set to 1.

The same general findings are obtained when using graded-relevance evaluation for the top 10 results (n@DCG@10), summative evaluation on the whole document ranking (MAP), and evaluation that only considers assessed documents (BPREF). These results are reported in Table 2: boosting the body field or alternatively considering the body and title fields equally deliver the highest retrieval effectiveness across collections and search tasks. An exception is represented by the BPREF values obtained on CLEF2016, suggesting that the findings in this collection my be influenced by a large number of unassessed documents.

In summary, these experiments have indicated that boosting the title field over other fields does not improve retrieval effectiveness, independently of the retrieval task at hand; instead, field weighting for the body field appears often to be more important than that of the title field for increasing retrieval effectiveness.

Next, we analyze the retrieval experiments at a query-by-query level, focusing on title and body weighting only, because they have shown more effect than other fields to determine the best retrieval effectiveness.

We first analyze when it is better to search on the title field only, and when on the body field only. Figure 3 summarizes the results of this analysis across all collections, based on p@10. Results from both general search and exploratory search exhibit similar trends: searching on body only is mostly better than searching on title only, or on a combination of title and body.

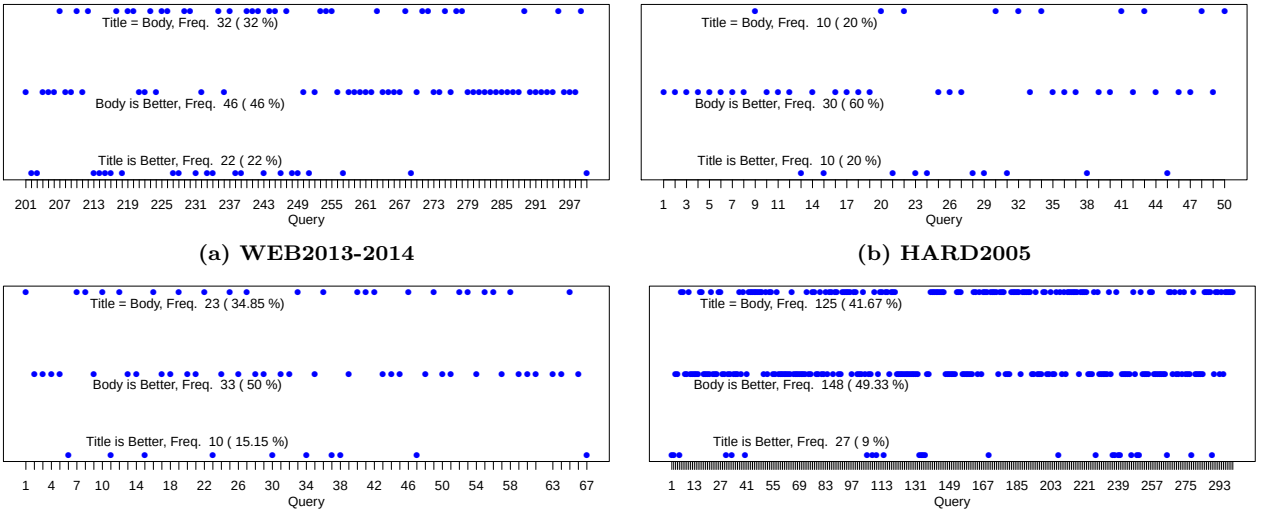
We further expand this analysis by considering four field weighting approaches: title only (title = 1), body only (body = 1), title weight equal to body (title = body), title weight more than body (title=3, body=1), and title weight less than body (title=1, body=3). This detailed query by query analysis allows us to determine whether differences are only affecting a small amount of queries, and whether the difference across weighting schema are substantial. Figure 5 summarizes these results for p@10. The results indicate that the body only strategy and the weighting the body field higher than the title field produce consistently better results on the CLEF collections, while on the WEB2013-2014 collections the strategies that ascribe more importance to the title are at par with those for the body. But, what queries then benefit by providing a higher weight to the title field?

4.3 Analysis of Navigational Queries

The results above have shown that while boosting the body field is generally likely to improve more the effectiveness than boosting the title field, there are queries for which boosting title does provide sensibly higher effectiveness. Specifically, we are interested to verify the intuition that if the query is navigational, then boosting the title field would provide better retrieval effectiveness than other weighting strategies. To this aim we focus on the WEB2013-2014 queries that have received navigational relevance assessments in the qrels: this is a subset of 10 queries[5].

Firstly, we are interested to understand the impact the different query natures have on the proportion of QIT in top ranked documents. Figure 4 compares the proportion of QIT for navigational queries against other queries. The proportion of QIT for navigational queries is only slightly higher than for non navigational queries for the top ten results. When we consider more results, the proportion of QIT for navigational queries drops rapidly and becomes less than the non-navigational queries. This is not surprising as the intent of navigational queries is to target specific documents. Documents relevant to a navigational query are likely to have titles that share most of the query terms. However,

[5]Query numbers 202, 223, 227, 257, 265, 266, 269, 273, 285, and 298



(a) WEB2013-2014 **(b) HARD2005**

(c) CLEF2015 **(d) CLEF2016**

Figure 3: Query-by-query analysis of when retrieving on body only is better than retrieving on title only.

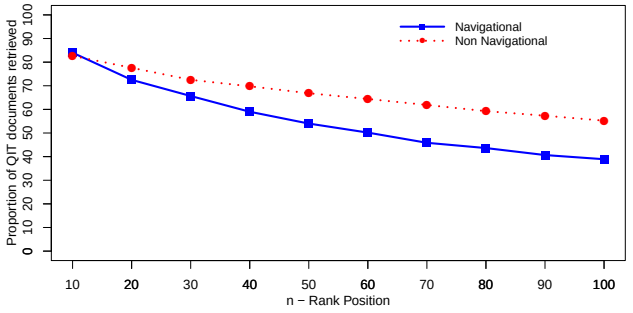


Figure 4: Proportion of QIT documents for navigational and non-navigational queries in WEB2013-2014

	Navigational	Other
Body > Title	0.2544	0.4815
Body Only	0.2775 (9.10%)	0.3841 (-20.22%)
Title Only	0.3919 (54.04%)	0.3320 (-31.05%)
Title = Body	0.2819 (10.83%)	0.4643 (-3.57%)
Title > Body	0.4032 (58.51 %)	0.4055 (-15.77%)

Table 3: Reciprocal rank (RR) values obtained for navigational queries VS other queries in WEB2013-2014 – percentage differences are calculated over the RR value obtained by weighting the body field higher than the query field. While the raw RR scores are not comparable, the relative differences can be compared. Results show that for navigational queries more gains are obtained when giving more importance to the title field; the opposite is found for the other queries. Differences for navigational queries are not statistically significant (10 queries only).

the number of documents explicitly targeted by navigational queries are limited. Therefore, the number of search results with QIT drops rapidly as we move to examine lower ranked documents.

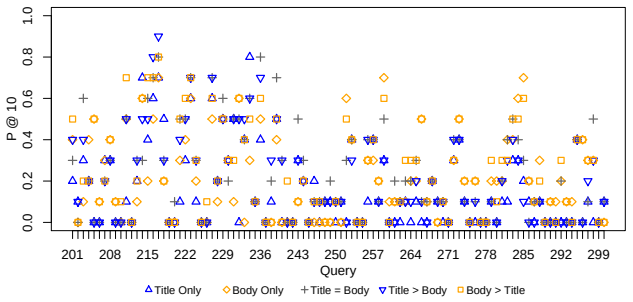
Secondly, we analyze navigational queries using the reciprocal rank measure (RR) and consider as relevant only those documents that have been assessed as navigational. We compare these results with the RR values obtained by the other queries in WEB2013-2014, but when considering as relevant all documents that have an assessment of at least partially relevant. We use reciprocal rank and these evaluation settings because for navigational intents it is likely that only one or a handful of documents are relevant to the query and the retrieval of that only document at a high rank position is of higher importance than the precision at a certain cutoff.

Table 3 reports the RR results for navigational queries VS the remaining queries for WEB2013-2014. While the raw values are not directly comparable because measured on different queries, the relative differences between the different weighting approaches are. The results show that for navigational queries, giving more importance to the title field over the body by boosting or by considering title only, lead to increased RR than boosting the body field. The contrary happens for non-navigational queries, where boosting the body field over the title lead to higher RR values than other weighting arrangements.

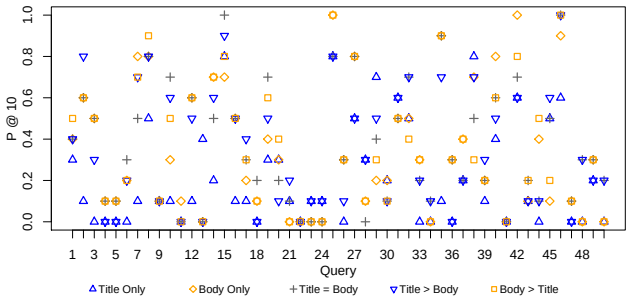
5. DISCUSSION

In this section we discuss the key findings from our experiments and possible implications.

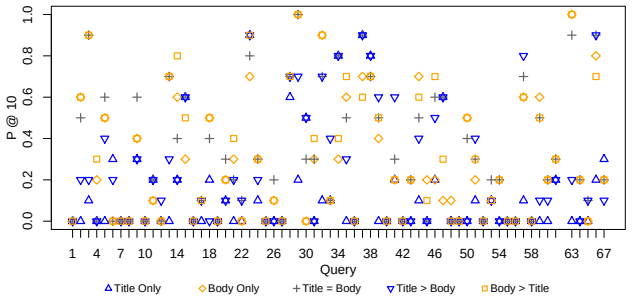
With regards to the presence of query terms in titles (RQ1), the results in Figure 1 confirm that top ranked documents are likely to contain *query-in-title* (QIT), i.e. query matches on the document field. However, as hypothesized, the QIT values for general web and newswire search tasks are significantly higher than for the more exploratory tasks as those represented by the consumer health search collections: in particular, most of the top ranked documents in CLEF2016 do not have QIT. Nevertheless, we also showed that the oc-



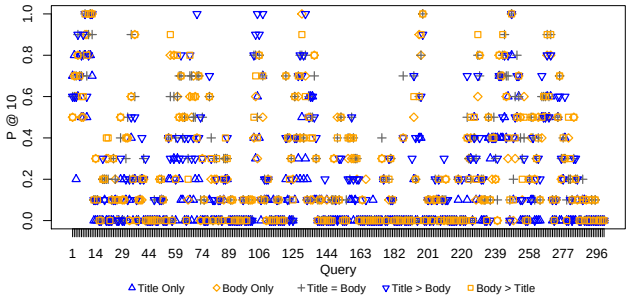
(a) WEB2013-2014



(b) HARD2005



(c) CLEF2015



(d) CLEF2016

Figure 5: Query-by-query analysis of p@10 results across different settings of title and body fields weighting.

currence of query terms in the titles is higher in relevant documents than in not-relevant ones across all collections.

With regards to the impact of field boosting on retrieval effectiveness (RQ2), we found that boosting the body field delivers better results than boosting other fields, and in particular title, see Figures 2c and 2d. This result is in contrast with common advice and the practice of boosting the title field over matches on the body field. And this is so for both exploratory web search, as one could intuitively hypothesize, and, somewhat unexpectedly, also for general web search. This result further demonstrates that field boosting and QIT ratios are not connected, as instead it was thought before [7].

Nevertheless, these results do not mean that the title field is not important. The results in Figures 2a, 2b, 2c and 2d show that the best p@10 values were obtained when the title field was included as a source of retrieval; this result is confirmed when using other evaluation measures (Table 2).

Furthermore, we found that boosting the title field, at least as high as the body field, does deliver increased retrieval effectiveness for navigational queries (Table 3). This means that the optimal field weight seems dependent more on the expected results (navigational VS ad-hoc or exploratory) than the actual tasks. An interesting direction for future work is the creation of adaptive methods that boost matches in specific fields according to the query, e.g., boost title if the query is navigational. Such adaptive method for field weighting has also been put forward by Trotman [21]. In Table 4 we report the optimal effectiveness (p@10) such an adaptive system would have if it was able to select the best field weighting on a per-query basis. The results are compared with the best effectiveness obtained in our experiments for methods that set a uniform field weight for all the queries in a specific collection (e.g., boosting title of a weight 1 and body of a weight 3 for the WEB2013-2014 collection). Results in Table 4 support Trotman's [21] findings that the creation of adaptive methods could significantly improve the performance of a retrieval system.

	Uniform	Adaptive	%Improv.
WEB2013-2014	0.2440	0.3050	25.00
HARD2005	0.3640	0.4580	25.82
CLEF2015	0.3136	0.3788	20.77
CLEF2016	0.2530	0.3107	22.79

Table 4: Potential effectiveness improvements (p@10) using an adaptive boosting approach. All differences between the uniform and the adaptive methods are statistically significant ($p < 0.01$).

6. LIMITATIONS AND CONCLUSIONS

In this paper, we conducted empirical experiments to challenge the common assumption that the title field of a web page should be boosted above other fields in field based retrieval methods.

First, we found that the proportion of top ranked documents with query in title (QIT) for general web search tasks is significantly higher than for exploratory search tasks such as consumer health search. This confirms our hypothesis that, because in exploratory search tasks queries tend to be circumlocutory and vague, query terms are unlikely to match documents' titles, which usually instead contain specific, topical keywords.

Second, our empirical results suggest that boosting the body field is better than boosting the title field of a web page to improve general retrieval effectiveness. The body field, in fact, is found to be more important for both general search tasks and for exploratory search tasks; proving

wrong the general advice of boosting matches in the title filed over other fields. Nevertheless, this does not mean that the title field is not important. Our experiments show that the best performance is gained by considering also the title, although not by boosting it. We also found that documents with QIT tend to be more likely relevant than documents without QIT. In contrast, our experiments with navigational queries show that, for these queries, the title field is indeed more important than the body field – and retrieval gain is obtained if matches in title are boosted.

Our experiments have a number of limitations. Firstly, we experimented with one retrieval system, Elasticsearch, and one field based model, BM25F. While this choice limits the generalizability of the findings, we note that Elasticsearch and BM25F are the most common solutions for enterprise and other small to medium search setups. Secondly, we only consider four main fields of web pages: title, meta, headers and body. We did not consider other fields such as anchor text information, which has been shown to be a valuable source of retrieval improvement, especially for navigational queries [4]. Lastly, we have not optimized the parameters of the BM25F function used for retrieval (i.e., B_f and K_1 value): the tuning of this parameter is not only important for increasing retrieval effectiveness, but it may also be required on a per-field base [18, 24].

Future work will address the limitations of our experiments, along with expanding this study to learning to rank settings [10] and exploring adaptive strategies to determine query-based field weighting.

Acknowledgment: Jimmy conducted this research as part of his doctoral study which is sponsored by Indonesia Endowment Fund for Education (Lembaga Pengelola Dana Pendidikan / LPDP).

7. REFERENCES

[1] J. Allan. HARD Track Overview in TREC 2005 High Accuracy Retrieval from Documents. In *Proceedings of TREC 2005*, 2005.

[2] C. L. A. Clarke, E. Agichtein, S. Dumais, and R. W. White. The Influence of Caption Features on Clickthrough Patterns in Web Search. In *Proceedings of ACM SIGIR 2007*, pages 135–142, New York, NY, USA, 2007. ACM.

[3] K. Collins-Thompson, C. Macdonald, P. Bennett, F. Diaz, and E. M. Voorhees. TREC 2014 web track overview. In *Proceedings of TREC 2014*, 2015.

[4] N. Craswell, D. Hawking, and S. Robertson. Effective Site Finding Using Link Anchor Information. In *Proceedings of ACM SIGIR 2001*, pages 250–257, New York, NY, USA, 2001.

[5] F. Diaz. Learning to Rank with Labeled Features. In *Proceedings of ACM ICTIR 2016*, pages 41–44, New York, NY, USA, 2016.

[6] I. S. Graham. *The HTML SourceBook*. John Wiley & Sons, Inc., New York, NY, USA, 1995.

[7] H. Joho, D. Hannah, and J. M. Jose. Emulating Query-biased Summaries Using Document Titles. In *Proceedings of ACM SIGIR 2008*, pages 709–710, New York, NY, USA, 2008.

[8] Kevyn Collins-Thompson, P. Bennett, C. Clarke, F. Diaz, and E. M. Voorhees. TREC 2013 Web Track Overview. In *Proceedings of TREC 2013*, 2013.

[9] J. Y. Kim and W. B. Croft. A Field Relevance Model for Structured Document Retrieval. In *Proceedings of ECIR 2012*, pages 97–108, Berlin, Heidelberg, 2012.

[10] T.-Y. Liu. Learning to Rank for Information Retrieval. *Foundation and Trends in Information Retrieval*, 3(3):225–331, 2009.

[11] W. Lu, S. Robertson, and A. MacFarlane. Field-weighted XML Retrieval Based on BM25. In *Proceedings of INEX 2005*, pages 161–171, Berlin, Heidelberg, 2006.

[12] C. Macdonald, R. McCreadie, R. L. T. Santos, and I. Ounis. From puppy to maturity: Experiences in developing terrier. *Open Source Information Retrieval*, 60, 2012.

[13] A. Molinari, G. Pasi, and R. A. M. Pereira. An Indexing Model of HTML Documents. In *Proceedings of ACM SAC 2003*, pages 834–840, New York, NY, USA, 2003.

[14] P. Ogilvie and J. Callan. Combining document representations for known-item search. In *Proceedings of ACM SIGIR 2003*, pages 143–150, 2003.

[15] J. Palotti, G. Zuccon, L. Goeuriot, L. Kelly, A. Hanbury, G. J. Jones, M. Lupu, and P. Pecina. CLEF eHealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. In *Proceeding of CLEF 2015*, 2015.

[16] J. Pérez-Iglesias, J. R. Pérez-Agüera, V. Fresno, and Y. Z. Feinstein. Integrating the probabilistic models BM25/BM25F into Lucene. *arXiv preprint arXiv:0911.5046*, 2009.

[17] S. Robertson. *The Probabilistic Relevance Framework: BM25 and Beyond*. Foundation and Trends in Information Retrieval, vol 3, no 4, 2009.

[18] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 Extension to Multiple Weighted Fields. In *Proceedings of ACM CIKM 2004*, pages 42–49, New York, NY, USA, 2004. ACM.

[19] D. Shin, H. Jang, and H. Jin. BUS: An Effective Indexing and Retrieval Scheme in Structured Documents. In *Proceedings of the ACM DL 1998*, pages 235–243, New York, NY, USA, 1998.

[20] I. Stanton, S. Ieong, and N. Mishra. Circumlocution in Diagnostic Medical Queries. In *Proceedings of ACM SIGIR 2014*, pages 133–142, New York, NY, USA, 2014.

[21] A. Trotman. Optimal Structure Weighted Retrieval. In *Proceeding of ADCS 2004*, 2004.

[22] R. Wilkinson. Effective Retrieval of Structured Documents. In *Proceedings of ACM SIGIR 1994*, SIGIR '94, pages 311–317, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[23] H. Xu, Z. Yang, B. Wang, B. Liu, J. Cheng, Y. Liu, Z. Yang, X. Cheng, and S. Bai. TREC 11 Experiments at CAS-ICT: Filtering and Web. In *Proceeding of TREC 2002*, 2002.

[24] H. Zaragoza, N. Craswell, M. J. Taylor, S. Saria, and S. E. Robertson. Microsoft Cambridge at TREC 13: Web and Hard Tracks. In *Proceeding of TREC 2004*, volume 4, page 1, 2004.

[25] G. Zuccon, B. Koopman, and J. Palotti. Diagnose this if you can. In *ECIR 2015 Workshop on Medical Information Retrieval (MedIR)*, pages 562–567, Vienna, Australia, 2015.

[26] G. Zuccon, J. Palotti, L. Goeuriot, L. Kelly, M. Lupu, P. Pecina, H. Mueller, J. Budaher, and A. Deacon. The IR Task at the CLEF eHealth evaluation lab 2016: user-centred health information retrieval. In *CLEF 2016 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS*, 2016.