Measuring the relative importance of full text sections for information retrieval from scientific literature.

Lana Yeganova, Won Kim, Donald C. Comeau, W. John Wilbur, and Zhiyong Lu
National Center for Biotechnology Information (NCBI),
National Library of Medicine (NLM), National Institutes of Health (NIH), USA

Abstract

With the growing availability of full-text articles, integrating abstracts and full texts of documents into a unified representation is essential for comprehensive search of scientific literature. However, previous studies have shown that naïvely merging abstracts with full texts of articles does not consistently yield better performance. Balancing the contribution of query terms appearing in the abstract and in sections of different importance in full text articles remains a challenge both with traditional bag-of-words IR approaches and for neural retrieval methods.

In this work we establish the connection between the BM25 score of a query term appearing in a section of a full text document and the probability of that document being clicked or identified as relevant. Probability is computed using Pool Adjacent Violators (PAV), an isotonic regression algorithm, providing a maximum likelihood estimate based on the observed data. Using this probabilistic transformation of BM25 scores we show an improved performance on the PubMed Click dataset developed and presented in this study, as well as on the 2007 TREC Genomics collection.

1 Introduction

PubMed (https://pubmed.gov) is a search engine providing access to a collection of more than 30 million biomedical abstracts. Of these, about 5 million have full text available in PubMed Central (PMC; https://www.ncbi.nlm.nih.gov/pmc). Millions of users search PubMed and PMC daily (Fiorini, Canese, et al., 2018). However, it is not currently possible for a user to simultaneously query the contents of both databases with a single integrated search.

With the growing availability of full-text articles, integrating these two rich resources to allow a unified retrieval becomes an essential goal, which has potential for improving information retrieval and the user search experience (Fiorini, Leaman, Lipman, & Lu, 2018). An obvious benefit is improving the handling of queries that produce limited or no retrieval in PubMed. In many instances, incorporating full text information can yield useful retrieval results. For example, the query *cd40 fmf* retrieves no articles in PubMed, but finds 60 articles in PMC discussing protein *cd40* and a computational technique of flow microfluorometry (FMF).

A number of studies have pointed out the benefits of full text for a range of text mining tasks (Cejuela et al., 2014; Cohen, Johnson, Verspoor, Roeder, & Hunter, 2010; J. Kim, Kim, Han, & Rebholz-Schuhmann, 2015; Westergaard, Stærfeldt, Tønsberg, Jensen , & Brunak, 2018) and demonstrated improved performance on named entity recognition, relation extraction, and other natural language processing tasks (Wei, Allot, Leaman, & Lu, 2019). For information retrieval, however, combining the full text of some papers with only the abstracts of others is not a trivial endeavor. Naïvely merging the body text of articles with abstract data, naturally increases the recall, but at a cost in precision, generally degrading the overall quality of the combined search (W. Kim, Yeganova, Comeau, Wilbur, & Lu, 2018; Jimmy Lin, 2009). This can be explained by several complexities associated with full texts, such as multiple subtopics often being discussed in a full-length article or information being mentioned in the form of conjecture or a proposal for future work. In addition, not every record matching the query is focused on the query subject, as query words may be mentioned in passing, which is more common in full text. Another challenge in incorporating full text in retrieval is merging sources of information with different characteristics: the abstract, generally a concise summary on the topic of the study, versus a lengthy detailed description provided in full text. To address that, recent studies have attempted to use full text in a more targeted way — by performing paragraph-level retrieval (Hersh, Cohen, Ruslen, & Roberts, 2007; Jimmy Lin, 2009), passage-level retrieval (Sarrouti & El Alaoui, 2017) or sentence-level retrieval (Allot et al., 2019; Blanco &

Proceedings of the BioNLP 2021 workshop, pages 247–256
June 11, 2021. ©2021 Association for Computational Linguistics

Zaragoza, 2010). LitSense (Allot et al., 2019), for example, searches over a half-billion sentences from the combined text of 30+ million PubMed records and ~3 million open access full-text articles in PMC.

Towards the overarching goal of improving PubMed document retrieval by incorporating the full texts of articles in PMC, in this work we lay the groundwork by studying strategies for integrating full text information with abstract for one query token at a time. We choose to use BM25, a classical term weighting approach, as a base token score. We, however, observe that token BM25 scores are not directly comparable between the sections of a full text article – the same BM25 score may have a different significance depending on the section. To address variable significance of sections, we propose converting BM25 section scores into probabilities of a document being clicked and using these probabilities to compute the overall token score. To summarize, given a single token in a query, we 1) define how to compute section scores, 2) examine the relative importance of different sections in the full text, and 3) study how to combine section scores from a document.

To examine these questions, we use two evaluation datasets. One is a standard TREC dataset frequently used for evaluating ad-hoc information retrieval. The second is a dataset we created based on PubMed user click information. The dataset is constructed from PubMed queries and clicks under the assumption that a clicked document is relevant to a user issuing a query. The dataset is used for both training and evaluation.

Neural retrieval models have been extensively studies in recent years in the context of Information Retrieval (Guo et al., 2020; Jimmy Lin et al., 2021). However, despite significant advances, they show no consistent improvement over traditional *bag of words* IR methods (Chen & Hersh, 2020; Zhang et al., 2020). BM25 remains in the core of most production search systems, including Lucene's search engine and PubMed. In addition, many relevance ranking algorithms rely on BM25 as a preliminary retrieval step, followed by re-ranking of the top scoring documents (Fiorini, Canese, et al., 2018).

In the next section, we describe the evaluation datasets, and lay out a retrieval framework for studying the problem at hand. Then, we describe our approach of converting the raw BM25 section score into the probability of document relevance. Such probabilities are comparable across the sections of full text documents, including the abstract. In section 4 we learn how to combine them in a way which accounts for the relative importance of sections. Results are presented in section 5, followed by the Discussion and Conclusions section.

2 Evaluation Datasets

Retrieval methods are generally evaluated based on how the retrieval output compares to a gold standard. A gold standard is a set of records judged for relevance to a query that provides a benchmark against which to measure the quality of search results. This approach is used at the annual Text Retrieval Conference (TREC), run by the National Institute of Standards and Technology (NIST) (Voorhees, 2001). NIST develops a list of queries, called topics, and provides large test collections and uniform scoring procedures. The difficulty with this approach is that a gold standard is created by human experts which makes the evaluation expensive, time consuming, and therefore not available for large scale experiments involving thousands of queries. To compare different retrieval approaches without a manually created gold standard we describe semi-automatically created test data based on indirect human judgements that can be utilized in our setting. The PubMed User Click dataset is created based on retrospective analysis of PubMed queries under the assumption that a clicked document is relevant to a user issuing a query. In our study we use both, the TREC 2007 Genomics and PubMed user click datasets.

TREC 2007 Genomics dataset. The Genomics dataset (Hersh et al., 2007) consists of 36 queries, called *topics*, and 162,259 full-text articles from Highwire Press (http:// highwire.stanford.edu/). 160K of these documents were successfully mapped to their corresponding PubMed Identifiers (PMIDs) and are the basis of our experiments. Each document is split into *legal spans* corresponding to paragraphs in the articles, amounting to over 12 million legal spans. For each of the 36 *topics* human relevance judgements are provided on the paragraph level. Following previous studies, a document is labeled positive, if it contains at least 1 paragraph judged to be relevant to the query.

The query topics are presented in the form of biological questions, such as:

What toxicities are associated with etidronate?

What signs or symptoms are caused by human parvovirus infection?

These question-like topic formulations contain generic words, that are not representative of the specific information need of a user, such as “what”, “associated”, etc. We applied a combination of frequency-based techniques and manual validation to filter these stop words out and used the remaining 165 content terms for our analysis.

PubMed Click Dataset. The dataset is constructed from PubMed queries and clicks, under the general assumption that a clicked document is relevant to a user issuing a query.

The presence of a query token in the title is known to present a strong signal associated with a document being clicked (W. Kim et al., 2018; Resnick, 1961). Users searching PubMed only see the title of the document on the DocSum page and not the abstract or the full text. If query tokens do not appear in the title, then predictions on the abstract or the full text can only be effective to the extent they predict something about the title that makes the user choose to click. This is a weaker signal and would be obscured by query words appearing in a title. To remove this bias, we only consider documents for which none of the query tokens appear in the title. Note that since the document is retrieved via PubMed, all query tokens must be found in the title, abstract or article citation information. We collect only retrieved documents for which none of the query tokens appear in the title and all of them appear in the abstract.

Clicked documents are assumed to be relevant to the user issuing the query, and we label a clicked document as a positive instance. We further assumed that documents displayed above the clicked document were seen by the user and rejected. These documents are labeled negative. Clicks on the top rank are ignored as a precaution, as those clicks might simply represent a user’s urge to click on something indiscriminately. Documents displayed below the lowest clicked document on the document summary page are ignored as the user may not have considered them.

The same query string may be searched multiple times within a period of time and subsequently may result in different articles displayed and different documents clicked. In addition, a query within a single search may receive multiple clicks on the same page. To account for these user search actions, we merge the data for the evaluation dataset as follows. Given a unique query string, we collect all positive and negative data points associated with each click instance, and remove from the negative set those documents that also appear as positives following the reasoning: if a document is thought to be relevant by at least one user we consider it relevant for that query string.

Using this dataset, for each query token we wish to compare its score coming from a document’s abstract versus the body text. First, to directly measure the benefit of full text, for each query in the PubMed Click Dataset, we perform this comparison on a subset of documents in the dataset that have full text available in PMC. Second, for each query in the dataset, we perform the comparison on all documents available in the PubMed Click dataset. This includes documents that do and do not have the full text available, as in production PubMed.

We randomly sampled 2 million unique queries from the PubMed query log in 2017, which retrieved at least one positive document. On average there are 6.60 documents collected for each query, an average of ~30% of which are labeled positive. Of 6.60 documents available for each query, only 2.65 documents have full text available in PubMed Central (~40%). We separated two thirds of queries for training PAV functions described in the next section, and one third for testing. 634,364 queries along with collected labeled documents comprise the test portion of the PubMed click dataset. A subset of that dataset that includes queries for which all retrieved documents have full text available constitutes 232,636 queries, and will be referred to as Set_FT.

3 Methods – Using Full Text to score a query token

Here we examine how to optimally use BM25 scores coming from the abstracts and full text paragraphs to improve retrieval performance. We first define the score of a token within a full-text section, which then we transform into a probability of that document being relevant given the score and the section. We then learn how to combine these section-based token scores into an overall score predicting the probability of a document being relevant.

3.1 Obtaining Full Text

We obtain full text documents from the PubMed Central full text collection in BioC (https://www.ncbi.nlm.nih.gov/research/bionlp/APIs) (Comeau, Wei, Doğan, & Z., 2019). This collection contains about 5 million full text manuscripts. BioC allows one to obtain full text information by paragraphs.

Full-text articles are typically comprised of sections presented in a logical sequence. Sections such as Introduction, Materials and Methods, Results, and Discussion predominantly appear as they represent the logical sequence in scientific writing. Frequently, however, sections carrying similar types of information are referred to differently depending on the journal, the requirements of the publishing entity, and author writing style. For example, Introduction and Background section titles are used interchangeably. Results sections can be also referred to as Results and Experiments, etc. Using BioC provided section type identifiers that are based on the labels and regular expressions found in (Kafkas et al., 2015). To normalize section titles, we concentrate on the following section types: Abstract, Abbreviation, Caption, Discussion, Case, Keyword, Conclusion, Result, Methods, Introduction, Generic Section Title, Supplement, and Appendix. In what follows, all the sections other than the Abstract text will be referred to as body sections or full text sections.

3.2 Defining the score of a token in a section

Given a token t we can compute a BM25 score s_t representing relevance of the token to a paragraph of text. The score is a product of the IDF weight and a local weighting factor that is zero if t does not occur in the paragraph. Using BM25 scoring of tokens in paragraphs, our goal is to devise a number representing the full text and its contribution to an overall document score that predicts user clicks based on each token in a query.

Since there are generally multiple paragraphs within each section of a paper, we keep the largest BM25 score for a token in a section paragraph and call it the BM25 score of the section type (*stype*) in a full text document and denote it s_t^{stype}. Keeping the maximum score is plausible because it is not affected by the size of the section (Jimmy Lin, 2009). Thus, given a token, for any document we have potentially thirteen different BM25 scores for that token, one from each section type.

Because of the structure of full text documents, the appearance of a token in different sections makes different contributions to the relevance of the document. The same BM25 score may have a different significance depending on the section. For example, a high score in the Results section would likely be more indicative of importance than if it occurred in the Methods section of a paper. To address the issue of variable significance of sections, we convert these BM25 section scores into probabilities of a document being clicked. The Pool Adjacent Violators (PAV) Algorithm (Ayer, Brunk, Ewing, Reid, & Silverman, 1954; Hardle, 1991; Wilbur, Yeganova, & Kim, 2005) is ideal for this purpose.

3.3 Training a PAV Function

Given a set of labeled data points along with their scores with the property that the higher the score the more likely a point is in the positive class, PAV is a simple and efficient algorithm that derives from such data a monotonically non-decreasing estimate of the probability that a point is in the positive class. Among non-decreasing functions that estimate the probability of a point being positive as a function of score, the PAV function assigns the highest likelihood to the actual observed class of the data points. Using training data, we apply PAV to the BM25 scores coming from each section type and obtain a function, p_{stype}, that predicts the probability of relevance. By nature of the monotonically non-decreasing estimate, the probabilities satisfy:

$$s_{t_i}^{stype} \le s_{t_j}^{stype} \Rightarrow p_{stype}\left(s_{t_i}^{stype}\right) \le p_{stype}\left(s_{t_j}^{stype}\right).$$

All scores from single tokens from queries appearing in training documents are distinct data points included for learning these PAV-derived probabilities. The stepwise linear PAV function for each of the thirteen document sections are presented in Figure 1. Results are presented in four blocks, each block comparing three body section PAV probability functions to the abstract probability function. The figures show that there is a difference between the sections in their relative importance. Given two sections, a higher BM25 token score from one section does not necessarily translate to a higher probability of relevance compared to the other section. If one section is more important for retrieval than the other, the same BM25 score in each section will lead to a higher probability in a more important section. Abrupt jumps may be due to sparseness of data This will have implications for retrieval.

The PAV-based probabilistic transformation allows one to directly compare the value of section scores to each other. A clear conclusion here is that the raw BM25 scores do not well reflect the relative importance of different body sections, as expected.

3.4 Combining Scores from Different Sections of the Body Text

Now we examine how to combine these probability scores coming from different sections into a single document score that predicts the document being relevant. Let us denote the probability of relevance given BM25 section scores as $p(rel|BM25 \text{ section scores})$. Then, the log odds ratio, defined as

$$\log\left[\frac{p(rel|BM25 \text{ section scores})}{p(\neg rel|BM25 \text{ section scores})}\right] \quad (1)$$

is monotonically related to the probability of relevance. We apply Bayes' Theorem.

$$\log\left[\frac{p(rel|BM25 \text{ section scores})}{p(\neg rel|BM25 \text{ section scores})}\right] \quad (2)$$
$$= \log\left[\frac{p(BM25 \text{ section scores}|rel)p(rel)}{p(BM25 \text{ section scores}|\neg rel)p(\neg rel)}\right]$$

The naïve Bayes' assumption will allow us to factor the right side of (2) as

$$\log\left[\frac{p(BM25 \text{ section scores}|rel)p(rel)}{p(BM25 \text{ secttion scores}|\neg rel)p(\neg \mathrm{rel})}\right] \quad (3)$$
$$= \log\left[\frac{\prod_{stype} p(s^{stype}|rel)}{\prod_{stype} p(s^{stype}|\neg rel)}\right] +$$
$$\log\left[\frac{p(rel)}{p(\neg rel)}\right].$$

The second term on the right in equation 3 is a constant and can be disregarded, as it will not affect the ranking. The first term on the right side of equation 3 can be rewritten as:

$$\log\left[\frac{\prod_{stype} p(s^{stype}|rel)}{\prod_{stype} p(s^{stype}|\neg rel)}\right] \quad (4)$$
$$= \sum_{stype} \log\left[\frac{p(rel|s^{stype})}{1-p(rel|s^{stype})}\Big/\frac{p(rel)}{1-p(rel)}\right].$$

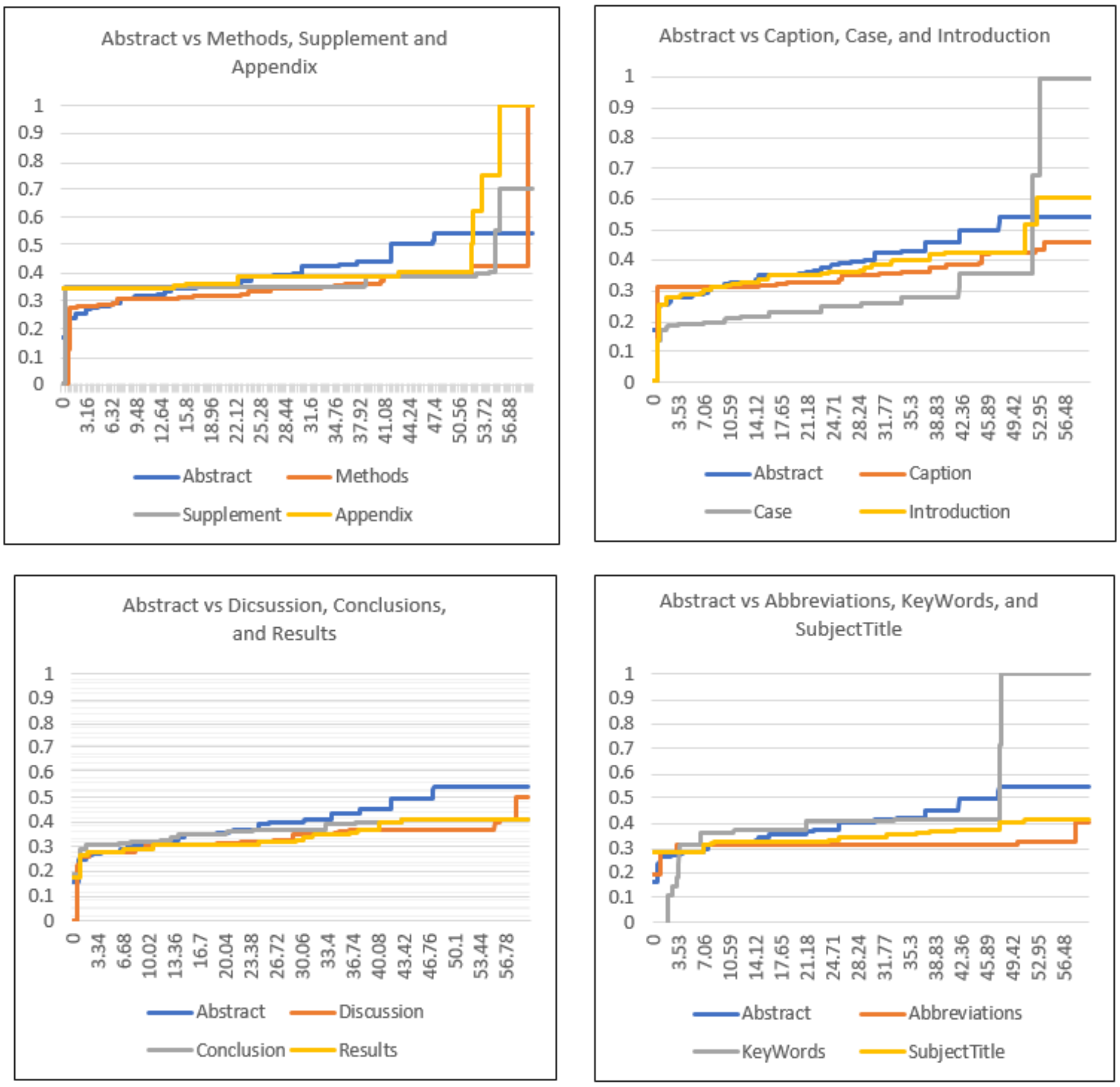
The right side of equation 4 is monotonically related to the left side of equation 2, and consequently should rank documents in the order of their probability of being relevant. This is the ideal ranking according to the probability ranking principle (Robertson, Walker, Jones, Hancock-Beaulieu, & Gatford, 1994). Here $p(rel|s^{stype}) = p_{stype}(s^{stype})$ is the PAV determined probability estimate for the section type, while $p(rel)$ is the fraction of positive documents in the training set. Based on these results we define the log odds score of a token in a section as

$$\log_{odds^{stype}(t)} = \log\left[\frac{p_{stype}(s_t^{stype})}{1-p_{stype}(s_t^{stype})}\Big/\frac{p_{random}}{1-p_{random}}\right]. \quad (5)$$

where $p_{random} = p(rel)$. Such scores for tokens can be added if the naive assumption of independence of the BM25 scores on which they are based is reasonably accurate.

Now we test different ways of combining scores of a token from different sections to derive a full-text score for the token. In (Jimmy Lin, 2009), the author found that computing the article score as the maximum score over all spans is superior to computing the score for an article as sum of scores over all spans. Spans in that work were paragraphs of full text documents from the TREC genomics collection, which consists of 36 topics (query questions) and manually annotated spans representing 2,477 full-text articles. In contrast, (Hearst & Plaunt, 1993) found that using the sum of scores over all spans in scoring a document produces a superior ranking when evaluated on a data set of 43 queries and 274 full text documents. Spans in (Hearst & Plaunt, 1993) are computed segments correlating with subtopics of a full text paper and are different from paragraphs.

Taking these references into consideration, we study and compare the *Sum* and *Max* scoring strategies using BM25 raw scores and log odds of BM25 scores. BM25 on Abstracts is also computed as it is used in the PubMed search system.



Abstract vs Methods, Supplement and Appendix

Abstract vs Caption, Case, and Introduction

Abstract vs Dicsussion, Conclusions, and Results

Abstract vs Abbreviations, KeyWords, and SubjectTitle

Fig 1. In these four graphs 12 PAV functions for 12 different body sections are compared to the abstract PAV function. X-axis represents the BM25 score across all four graphs, Y axis represents the Probability of a click based on the section term score.

Sum LogOdds: The score of token *t* in document *d* is computed as the sum of log odds scores, as defined in (5), coming from sections within the full text document:

$$sco_{Sum\_LogOdds}(d,t) = \sum_{stype\in d} \log\_odds^{stype}(t) \qquad (6)$$

Max LogOdds: The score of token *t* in document is computed as the maximum log odds score coming from sections within the full text document:

$$sco_{Max\_LogOdds}(d,t) = \max\{\log\_odds^{stype}(t) | stype \in d\} \qquad (7)$$

Abstract BM25: The score of token *t* in document *d* is computed as the raw BM25 token score of the abstract

$$sco_{BM25\_Abs}(d,t) = s_t^{abs} \qquad (8)$$

Sum BM25: The score of token *t* in document *d* is computed as the sum of BM25 section token scores within the full text document

$$sco_{Sum\_BM25}(d,t) = \sum_{stype\in d} s_t^{stype} \qquad (9)$$

Max BM25: The score of token *t* in document *d* is computed as the highest BM25 section token score within the full text document

$$sco_{Max\_BM25}(d,t) = max_{stype \in d}\{s_t^{stype}\} \quad (10)$$

After trying scoring based directly on log odds using formulas (6) and (7), it was evident that we are dealing with two kinds of documents, which behave differently. Those documents that contain the search token only in the abstract receive a single score from the abstract, and *Sum* and *Max* really don't play a role. But for those documents having the term in multiple sections, *Sum* and *Max* do play a role, and the log odds scores are higher. In order to balance the scores for best results, we found it necessary to create PAV curves for *Sum* and *Max* scores just on documents with multiple sections providing scores. We simply use the probabilities based on a PAV curve for each type of document to rank the different types in the same ranking for retrieval. In what follows, we will continue to use the term LogOdds to refer to this scoring.

4 Results

Proposed methods are tested on the PubMed Click Dataset and on the TREC Genomics collection (Hersh et al., 2007).

4.1 The PubMed Click Dataset

To directly measure the benefit of full text, for each query in the PubMed Click Dataset we first compare the proposed scoring techniques on Set_FT. Set_FT is a subset of the PubMed Click dataset that includes queries for which all labeled documents in the evaluation dataset have full text available. Second, we extend this analysis to the whole test portion of the PubMed Click dataset. It contains queries and labeled documents, which may or may not have full text available. For each query token, we score its corresponding retrieved documents in the evaluation dataset and compute the average Precision using labels in the evaluation dataset. These are averaged over all tokens in a query, and then average over all queries producing the MAP results presented in Figure 2.

Figure 2 demonstrates our findings computed on the complete set of tokens available in the two test sets. We observe that the LogOdds probabilistic scoring approach significantly outperforms the BM25 scoring for both *Sum* and *Max* variants for the PubMed click data and Set_FT. A bigger difference is observed on Set_FT, where full text is available for every participating document. Additionally, we observe that LogOdds Sum computed on article full text outperforms the abstract score and the difference although small is statistically significant.

We conducted pairwise statistical tests for all methods to verify if the differences in performance for each pair of tests is significant. We used the "Percentile bootstrap" test at the 5% significance level which works well for our study because the distribution is symmetric around the MAP value (https://en.wikipedia.org/wiki/Bootstrapping). Differences between all pairs of methods are statistically significant, except for the Max LogOdds and the Abs BM25 for the Set_FT subset of PubMed Click Dataset.

Based on these results we believe that log odds scoring is a useful approach for retrieval incorporating body text. The intuition behind it is that BM25 scores have a different meaning depending on the sections from which they are derived as illustrated in Fig 1. For a single query token, results in Figure 2 also suggest that the *Sum* scoring approach provides a better estimate of token importance than the *Max* scoring approach when using the log odds scoring for the Click dataset. If sections within a full text document were truly independent from each other, Sum LogOdds would be the ideal method to score a single query token over the multiple sections in a document.

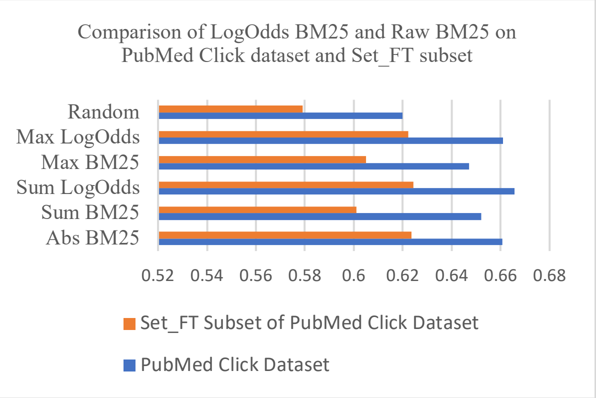


Figure 2. Average Precision for all query tokens is computed, averaged for each query and then over all queries for the PubMed Click dataset and its subset Set_FT. For both datasets, LogOdds Sum and LogOdds Max scoring methods demonstrate a significantly improved performance compared to Sum and Max on raw BM25 scores.

4.2 The TREC Genomics Dataset

We apply the proposed methods to each query token in the TREC dataset. We score the retrieved documents in the evaluation dataset and compute the Average Precision using gold standard labels. These are then averaged over all query tokens, and the MAP results are presented in Figure 3. Leave-one-out training strategy was used for each topic.

Figure 3 demonstrates our findings computed on non-stop word query tokens in the TREC Genomics Dataset. We observe that the Sum LogOdds probabilistic scoring significantly outperforms Sum BM25 scoring. Similarly, the Max LogOdds probabilistic scoring significantly outperforms Max BM25 scoring. Similar to the PubMed Click Dataset, here we observe that Sum LogOdds has a slight advantage over Max LogOdds, and both are competitive with the abstract BM25 score.

We conducted Wilcoxon signed rank test (https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test) at 5% significance level to verify if the differences in performance for each pair of tests is significant. The differences between Max LogOdds and Abs BM25 as well as Sum LogOdds and Abs BM25 are not statistically significant. The differences between all other pairs of methods are statistically significant.

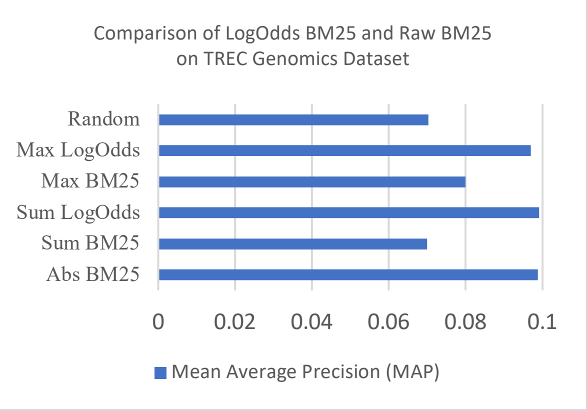


Figure 3. Mean Average Precision on TREC Genomics Dataset is computed on single tokens and averaged for all tokens in the experiment. Sum LogOdds and Max LogOdds demonstrate a significantly improved performance compared to those on raw BM25 scores.

5 Conclusions and Discussion

Based on the PubMed Click dataset and the TREC genomics dataset, we studied how to integrate full text and abstract information for scoring a query token. The main contribution of this work is to study the benefits of log odds of BM25 compared to raw BM25 scores. Our experimental results on both datasets support these important conclusions:

1. PAV based log odds scoring is a useful way to compare the contribution of a token in different sections of a document for predicting clicks. BM25 scores are not directly comparable with each other for making such predictions. The same BM25 score is of different value depending on the section type in which it is found.

2. We proposed two methods to compute the log odds body score by taking the sum or max of scores. In both cases, PAV based LogOdds scoring is significantly better than ranking based on raw BM25 scores. The difference between *Sum* and *Max* scoring is small.

For the PubMed Click dataset, using the Sum LogOdds score from the whole document for a query token produces better results than using only the abstract score. In the TREC genomics dataset, the performance of full text LogOdds is comparable to abstract only score. This is an important contribution and meaningful building block towards improving full text retrieval in PubMed. Our immediate plan is to extend this single token analysis to full queries.

6 Conclusions and Discussion

Based on the PubMed Click dataset and the TREC genomics dataset, we studied how to integrate full text and abstract information for scoring a query token. The main contribution of this work is to study the benefits of log odds of BM25 compared to raw BM25 scores. Our experimental results on both datasets support these important conclusions:

1. PAV based log odds scoring is a useful way to compare the contribution of a token in different sections of a document for predicting clicks. BM25 scores are not directly comparable with each other for making such predictions. The same BM25 score is of different value depending on the section type in which it is found.

2. We proposed two methods to compute the log odds body score by taking the sum or max of scores. In both cases, PAV based LogOdds scoring is significantly better than ranking based on raw BM25 scores. The difference between *Sum* and *Max* scoring is small.

For the PubMed Click dataset, using the Sum LogOdds score from the whole document for a

query token produces better results than using only the abstract score. In the TREC genomics dataset, the performance of full text LogOdds is comparable to abstract only score. This is an important contribution and meaningful building block towards improving full text retrieval in PubMed. Our immediate plan is to extend this single token analysis to full queries.

FUNDING
This work was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

References

Allot, A., Chen, Q., Kim, S., Vera Alvarez, R., Comeau, D. C., Wilbur, W. J., & Lu, Z. (2019). LitSense: making sense of biomedical literature at sentence level. *Nucleic Acids Research, 47*(Web Server issue).

Ayer, M., Brunk, H., Ewing, G., Reid, W., & Silverman, E. (1954). An empirical distribution function for sampling with incomplete information. *Ann Math Stat, 26*, 641-647.

Blanco, R., & Zaragoza, H. (2010). Finding Support Sentences for Entities. *SIGIR '10 Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*

Cejuela, J., McQuilton, P., Ponting, L., Marygold, S., Stefancsik, R., Millburn, G., . . . Consortium, F. (2014). tagtog: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles. *Database (Oxford)*. doi:10.1093/database/bau033. Print 2014.

Chen, J., & Hersh, W. R. (2020). A Comparative Analysis of System Features Used in the TREC-COVID Information Retrieval Challenge (Publication no. https://doi.org/10.1101/2020.10.15.20213645).

Cohen, K. B., Johnson, H., Verspoor, K., Roeder, C., & Hunter, L. (2010). The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics, 11*(492).

Comeau, D. C., Wei, C.-H., Doğan, R. I., & Z., L. (2019). PMC text mining subset in BioC: about 3 million full text articles and growing. *Bioinformatics, doi:10.1093/bioinformatics/btz070.*

Fiorini, N., Canese, K., Starchenko, G., Kireev, E., Kim, W., Miller, V., . . . Lu, Z. (2018). Best Match: New relevance search for PubMed. *PLOS Biology, 16*(8). doi:doi: 10.1371/journal.pbio.2005343

Fiorini, N., Leaman, R., Lipman, D. J., & Lu, Z. (2018). How user intelligence is improving PubMed. *Nature Biotechnology, 36*, 937–945.

Guo, J., Fan, Y., Pang, L., Yang, L., Ai, Q., Zamani, H., . . . Cheng, X. (2020). A Deep Look into Neural Ranking Models for Information Retrieval. *Journal of Information Processing and Management, 57*(6).

Hardle, W. (1991). *Smoothing techniques: with implementation in S*. New York: Springer-Verlag.

Hearst, M. A., & Plaunt, C. (1993). *Subtopic structuring for full-length document access*. Paper presented at the SIGIR93: 16th International ACM/SIGIR '93 Conference on Research and Development in Information Retrieval, Pittsburgh PA USA.

Hersh, W., Cohen, A., Ruslen, L., & Roberts, P. (2007). TREC 2007 Genomics Track Overview *Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007)*.

Kafkas, Ş., Pi, X., Marinos, N., Talo', F., Morrison, A., & McEntyre, J. R. (2015). Section level search functionality in Europe PMC. *Journal of Biomed Semantics*. doi:doi: 10.1186/s13326-015-0003-7

Kim, J., Kim, J., Han, X., & Rebholz-Schuhmann, D. (2015). Extending the evaluation of Genia Event task toward knowledge base construction and comparison to Gene Regulation Ontology task. *BMC Bioinformatics*. doi:10.1186/1471-2105-16-S10-S3. Epub 2015 Jul 13.

Kim, W., Yeganova, L., Comeau, D. C., Wilbur, W. J., & Lu, Z. (2018). MeSH-based dataset for measuring the relevance of text retrieval. *Proceedings of the BioNLP 2018 workshop*.

Lin, J. (2009). Is searching full text more effective than searching abstracts? *BMC Bioinformatics, 10*(46).

Lin, J., Ma, X., Lin, S.-C., Yang, J.-H., Pradeep, R., & Nogueira, R. (2021). Pyserini: An Easy-to-Use Python Toolkit to Support Replicable IR Research with Sparse and Dense Representations.

Resnick, A. (1961). Relative effectiveness of document titles and abstracts for determining relevance of documents. *Science, 134*(3484), pp. 1004–1006.

Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., & Gatford, M. (1994). Okapi at TREC-3. *Proceedings of the 3rd Text REtrieval Conference (TREC-3)*.

Sarrouti, M., & El Alaoui, S. O. (2017). A passage retrieval method based on probabilistic information retrieval model and UMLS concepts in biomedical question answering. . *Journal of Biomedical Informatics, 68*.

Voorhees, E. (2001). The philosophy of information retrieval evaluation. *CLEF 2001: Evaluation of Cross-Language Information Retrieval Systems, Volume 2406*, pp. 355–370.

Wei, C.-H., Allot, A., Leaman, R., & Lu, Z. (2019). PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research, 47*(W1).

Westergaard, D., Stærfeldt, H.-H., Tønsberg, C., Jensen , L. J., & Brunak, S. (2018). A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *Plos Computational Biology, 14*(2).

Wilbur, W. J., Yeganova, L., & Kim, W. (2005). The Synergy Between PAV and AdaBoost. *Machine Learning, 61*, 71-103.

Zhang, E., Gupta, N., Tang, R., Han, X., Pradeep, R., Lu, K., . . . Lin, J. (2020). Covidex: Neural Ranking Models and Keyword Search Infrastructure for the COVID-19 Open Research Dataset. *Proceedings of the First Workshop on Scholarly Document Processing*.