# Investigation of Passage Based Ranking Models to Improve Document Retrieval

Ghulam Sarwar[1(✉)], Colm O'Riordan[1], and John Newell[2]

[1] Department of Information Technology, National University of Ireland, Galway, Ireland
{g.sarwar1,colm.oriordan}@nuigalway.ie
[2] School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Ireland
john.newell@nuigalway.ie

**Abstract.** Passage retrieval deals with identifying and retrieving small but explanatory portions of a document that answers a user's query. In this paper, we focus on improving the document ranking by using different passage based evidence. Several similarity measures were evaluated and a more in-depth analysis was undertaken into the effect of varying specific. We have also explored the notion of query difficulty to understand whether the best performing passage-based approach helps to improve, or not, the performance of certain queries. Experimental results indicate that for the passage level technique, the worst-performing queries are damaged slightly and the those that perform well are boosted for the WebAp collection. However, our rank-based similarity function boosted the performance of the difficult queries in the Ohsumed collection.

**Keywords:** Document retrieval · Passage-based document retrieval · Passage similarity functions · Inverse rank · Query difficulty

## 1 Introduction

Information Retrieval (IR) deals with the organization, representation, and the retrieval of information from a large set of text documents. The retrieval of relevant information from large collections is a difficult problem; search queries and documents are typically expressed in natural language which introduces many problems such as ambiguity caused by the presence of synonyms and abbreviations, and issues arising from the *vocabulary difference problem* which occurs when the user expresses their information need with terms different to those used to express the same concept in the document collection.

Several models have been shown to be very effective in ranking documents in terms of their relevance to a user's query. The user formulates the query by expressing their information need in natural language. Approaches include different mathematical frameworks (vector space model, probabilistic models)


A. Fred et al. (Eds.): IC3K 2017, CCIS 976, pp. 100–117, 2019.
https://doi.org/10.1007/978-3-030-15640-4_6

to represent documents and queries and to formulate a comparison approach. The BM25 weighting scheme [1] derived within a probabilistic framework is a well-known effective one in estimating the relevance of a document to a query. The main goal of an IR system is to estimate the relevance of a document to a query; this notion of 'relevance' is often interpreted as measuring the level of similarity between a document to a query.

In IR, the traditional approaches consider the document as a single entity. However, some researchers choose to split the document into a separate passages given the intuition that a highly relevant passage may exist in a larger document which itself will be considered as non relevant. If a passage is indexed as an individual *pseudo-document*, the number of documents stored and indexed will increase significantly and in a result, it will effect the speed and cost of retrieval [2]. However, one may now retrieve relevant passages that occur in documents deemed not very relevant. Moreover, if the document returned as relevant is too long, it can be difficult for the users to find the appropriate relevant passages in the document. In other words, returning a large relevant document, while useful, still, puts an onus on the user to find the relevant passages. Therefore, we opt for the passage level retrieval approach to finding the relevant passage and aim to use that to improve the document ranking. The intuition behind our approach is that by identifying very relevant passages in a document we can better estimate the relevance of the overall document.

One can imagine the passages themselves as documents at indexing time. The division of these passages can be done in a number of ways. For example, either via some textual identifier e.g. paragraph markings (<$p$>), new line feed (/n) etc. or it can be defined by a number of words. A passage could be a sentence, a number of sentences or a paragraph itself. The passages can be considered as discrete passages with no intersection or can be viewed as overlapping passages.

In this paper, we extend our previous work [3] in which we utilised the inverse rank of a passage as a measure and compared it with the other passage based approaches. We recap some of our previous findings including the passage based equations and several figures and a table (Figures 1, 2(a), (b), Table 1 etc.) to support our new approaches and analysis undertaken in this paper. Our main goal is to generate new document rankings by computing the passage similarity and using this score (or its combination with document level similarity score) as a means to rank the overall document. In this extended work, we present a more extensive analysis of the SF2 approach (explained in Sect. 3) and also highlight the impact of difficult queries on the overall performance by analysing the ranking functions we used in our previous paper.

The main focus of our work is to see how effectively the passage level evidence affected the document retrieval. Furthermore, we extend our focus by doing a more in-depth analysis of the passage based functions based on different parameters and examining whether the difficult queries damages the passage based results or improves it. Factors such as different means to define passage boundaries are not of huge concern to us as present.

We have used the WebAp (Web Answer Passage)[1] test collection which is obtained from the 2004 TREC Terabyte Track Gov2 collection and the Ohsumed test collection [4] which comprises titles and/or abstracts from 270 Medline reference medical journals. The results show that different similarity functions behave differently across the two test collections. Furthermore, difficult queries have different characteristics and the impact on the overall performance.

The paper outline is as follows: Section 2 presents a brief overview of the previous work in passage level retrieval and some work done in query difficulty area. Section 3 gives an overview of the methodology employed, outlining the details of different similarity functions, the passage boundary approach, and the evaluation measures adopted in the experiments. Section 4 presents a brief explanation of the test collections used in the experiments and the assumptions made for them. Section 5 discusses different experimental results obtained. In Sect. 5.1, further analysis for the SF2 approach is presented. A discussion on the impact of difficult queries by using the passage level measures is presented in Sect. 5.2. Finally, Sect. 6 provides a summary of the main conclusions and outlines future work.

## 2 Related Work

In previous research, passage level retrieval has been studied in information retrieval from different perspectives. For defining the passage boundaries, several approaches have been used. Bounded passages, overlapping window size, text-tiling, usage of language models and arbitrary passages [5–9] are among the few main techniques. Window size approaches consider the word count to separate the passages from each other, irrespective of the written structure of the document. Overlapping window size is shown to be more effective and useful for the document retrieval [5]. Similarly, a variant of the same approach was used by Croft [10].

Jong [11] proposed an approach which involved considering the score of passages generated from an evaluation function to effectively retrieve documents in a Question Answering system. Their evaluation function calculates the proximity of the different terms used in the query with different passages and takes the maximum proximity score for the document ranking.

Callan [5] demonstrated that ordering documents based on the score of the best passage may be up to 20% more effective than standard document ranking. Similarly, for certain test collections, it was concluded that combining the document score with the best passage score gives improved results [12]. Buckley et al. also use the combination of both scores in a more complex manner, to generate scores for ranking [13]. Moreover, Hearst et al. [14] showed that instead of only using the best passage with the maximum score, adding other passages gives better overall ranking as compare to the ad-hoc document ranking approach.

Salton [15] discussed another idea to calculate the similarity of the passage to the query. They re-ranked and filtered out the documents that has a low passage

[1] https://ciir.cs.umass.edu/downloads/WebAP/.

score associated with it. They included all the passages that have a higher score than its overall document score, and then used these scores to raise, or lower, the final document rank. In this way, the document that has a lower score to the document level score but a higher score at passage level for certain passages, will get a better ranking score in the end.

Different language modelling approaches at passage level and document level have been used in the past to improve the document ranking [10,16]. A similar approach has been used by Bendersky et al. [7], where they used the measure of the document homogeneity and heterogeneity to combine the document and passage similarity with the query to retrieve the best documents. To use the passage level evidence, their scoring method used the maximum query-similarity score that is assigned to any passage in the document ranking. As for their passage based language model, they used the simple unigram based standard to estimate the probabilities at passage and document level. Moreover, Krikon and Kurland [17,18] used a different language modeling approach where they tried to improve the initial ranking of the documents by considering the centrality of the documents and the passages by building their respective graphs. The edges denote the inter-term similarities and the centrality is computed using the page rank approach. They reported that their approach performed better than the normal maximum passage approach and some variation of interpolation score of maximum passage score with document score.

Due to the recent improvements in the learning based models, researcher are using neural networks for passage evidences that could improve the Ad-hoc Document retrieval. Ai and Croft [19] developed a passage based neural model that uses the evidences given from the passages for the document retrieval. They used a learning based approach to weights the passages of different sizes and granularities and did not adopt the usual single window for passage extraction. They introduced a fusion framework that aggregates the passage score based on the its document properties and relation with the query characteristics. They compared their results with the work done by Liu et al. and Ponte et al. [10,20] and showed that their neural passage model out performed the previous passage based retrieval models. Similarly, Galko et al. [21] used the neural network based approach to improve the passage retrieval for the Question Answering (QA) task in the Biomedical Domain. They used the weighted combination of word embedding terms by using word2vec [22] and measured the consine distances between the query terms and the passages. Finally they compare their results with the previous neural-net based models and reported improvements with their approach.

In previous research it has been identified that a particular search approach may vary considerably in it performance across different queries. There are many potential underlying problems that may cause this: variation in query quality (specific, unambiguous queries through to vague ambiguous ones), nature of the document set, aspects of the weighting scheme or preprocessing approaches. Rather than merely considering the mean average precision, it can be very informative to consider the performance of individual queries. Identifying difficult

queries (those that the IR system produces low quality results) which could be the cause of decline MAP is an interesting problem. Mothe et al. [23] attempted to use the linguistic approaches in order to find the main reason of certain queries to become difficult. For each word, they computed the morphosyntactic category based on lexicons and a language model. They also calculated the semantic and syntactical features of each word by using wordnet and other analyzers. Looking at all these features they reported their correlation with the query difficulty. Similarly, He et al. [24] used a coherence-based approaches to measure the query difficulty. Their query based coherence scored illustrated the association with the average precision and they argued that this score can be used to anticipate the query difficulty.

## 3 Methodology

In traditional adhoc IR, a 'bag of words' model is adopted with no attention paid to word order or word position within a document. Weights are typically assigned to terms according to some heuristics, probability calculations or language model.

In this work, we view every document as being represented as passages or 'pseudo-documents' i.e. $d^{'} = \{p_1, p_2, \ldots p_n\}$. We attempt to better estimate $sim(d, q)$ by estimating $sim(d^{'}, q)$. Different similarity functions are designed in a way that different characteristics of the passage level results can be used alone, or in combination with the document level results. We define $sim(d^{'}, q)$ as $f(sim(p_i, q), sim(d, q))$.

### 3.1 Similarity Functions

Following is a brief description of these similarity functions in which different characteristics were computed from the passage level evidence:

- {**SF1**} Max Passage: One way to compute the $sim(d^{'}, q)$ is to consider the similarity and ranking of the passage that has the highest similarity score to the query as a representative of the similarity of the document.

$$sim(d^{'}, q) = max(sim(p_i, q))$$

- {**SF2**} Sum of passages: It is similar to the max passage approach, but instead of taking only the top passage, the top $k$ of the passages are taken and their similarity scores are combined by adding them together.

$$sim(d^{'}, q) = \sum_{i=1}^{k} [sim(p_i, q)]$$

– {**SF3**} Combination of document and passage similarity scores: In this case, the passage and document scores are combined and then the results are re-ranked based on the new score.

$$sim(d^{'}, q) = \alpha(max(sim(p_i, q))) + \beta(sim(d, q))$$

– {**SF4**} Inverse of rank: Rather than using the document or passage scores, the rank at which these passages are returned can also be used to find the similarity between the passages and the query. This can be calculated as follows:

$$sim(d^{'}, q) = (\frac{\sum_i \frac{1}{rankP_i}}{\#ofp_i}) \quad |p_i \in d^{'}$$

– {**SF5**} Weighted Inverse of Rank: Another way to take the rank of these passages into account is to take the sum of the inverse ranks and pay less attention to lower ranks. Hence, the higher ranks will impact more on the results as compare to the lower values and will effect the overall ranking.

$$sim(d^{'}, q) = \sum\nolimits_i (\frac{1}{rankP_i})^{\alpha} \quad |p_i \in d^{'}, \alpha > 1$$

### 3.2 Passage Boundaries

To run the experiments, all the documents and passages were first indexed in our IR system. We have used Solr 5.2.1[2] as a baseline system which is a high performance search server built using Apache Lucene Core. In this system, a vector space model is adopted with a weighting scheme based on the variation of tf-idf and Boolean model (BM) [25] is used.

We use two different test collections in the experiments. The WebAP test collection contains 6399 document and 150 queries in its dataset. We adopt overlapping windows for this collection and decompose each document into passages of length 250 words. This results in the creation of 140,000 passages for the WebAP collection. The second collection, the Ohsumed dataset, comprises 348,566 Medline abstracts as documents with 106 search queries. Given the relatively small document lengths, in defining passage boundaries, an overlapping window size of 30 words is used for this collection which creates a document set of passages of size 1.4 million pseudo-documents that gives 4–5 passages per document. We choose the half overlapping, fixed length window-size to index the documents, because these passages are more suitable computationally, convenient to use, and were proved to be very effective for document retrieval [5,10].

### 3.3 Evaluation

To evaluate the results and measure the quality of our approach, mean average precision (MAP) and precision@k are used as the evaluation metrics. The MAP

[2] http://lucene.apache.org/solr/5_2_1/index.html.

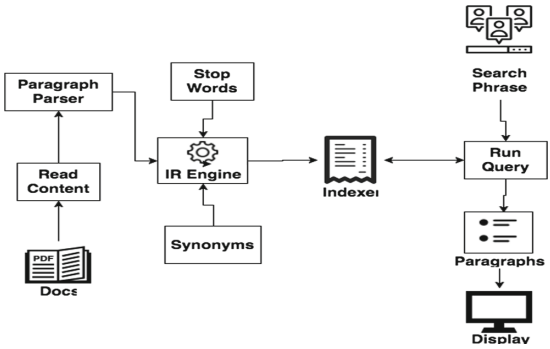



**Fig. 1.** Architectural diagram, extracted from [12].

value is used to give an overall view of the performance of the system with different similarity functions. Furthermore, precision@k was helpful in illustrating the behavior of the system with respect to correctly ranking relevant documents in the first $k$ positions.

## 4 Experimental Setup

In this section, we present a brief explanation of the test collections we used, and also some detail of different parameters that we consider in our experiments. Lastly we will describe the brief overview of the evaluation measures that we used in the experiments.

**Table 1.** MAP(%) for WebAp and Ohsumed collection at k = 5 and k = 10, extracted from [12].

| Similarity functions | MAP@5 (WebAP) | MAP@10 (WebAP) | MAP@5 (Ohsumed) | Map@10 (Ohsumed) |
|---|---|---|---|---|
| Document level (D) | 9.52 | 18.60 | 2.97 | 4.75 |
| Max passage (SF1) | 9.43 | 18.56 | 3.23 | 4.96 |
| Sum of passages (SF2) | 9.42 | 18.54 | 3.19 | **4.99** |
| Inverse of rank (SF4) | 9.42 | 18.56 | **3.27** | 4.89 |
| Weighted inverse of rank (SF5) | 9.43 | 18.58 | 3.20 | 4.98 |
| D+SF1 | 9.53 | 18.65 | 3.01 | 4.90 |
| D+SF2 | 9.53 | **18.67** | 2.82 | 4.74 |
| D+SF4 | 9.54 | 18.66 | 2.88 | 4.80 |
| D+SF5 | **9.55** | **18.67** | 2.80 | 4.60 |

### 4.1 Test Collections

For our experiments we used the two different test collections that are freely available to use for experimental purposes.The following is a brief explanation of both datasets.

**WebAp.** Web Answer Passage (WebAP) is a test collection, which is obtained from the 2004 TREC Terabyte Track Gov2 collection. The dataset contains 6399 documents and 150 query topics and relevance judgment of top 50 documents per query topic. It is created mainly for the purpose of evaluating passage level retrieval results [26] but has been used in question answering (QA) task to retrieve sentence level answers as well [27,28]. The query topic section contains keyword based queries and the normal queries. We generated the results against both types and here we reported the performances that are based on the keyword based queries. On average, these results performed overall 2% better than the normal query ones across all similarity functions. Annotation at passage level (GOOD, FAIR, PERFECT etc.) is also included in this test collection that can be used to differentiate the different passages in term of their relevance to the query. The annotators found 8027 relevant answer passages to 82 TREC queries, which is 97 passages per query on average. From these annotated passages, 43% of them are perfect answers, 44% are excellent, 10% are good and the rest are fair answers. We have saved these passage annotations while indexing them in the system, but, we have not used them in our evaluation criteria. As the size of all the documents are fairly large compare to the other test collections we came across, therefore, we divided passages using overlapping window based approach of size 250 words.

**Ohsumed.** The Ohsumed collection consists of titles and abstracts from 270 Medline reference medical journals. It contains 348,566 articles along with 106 search queries. In total, there are 16,140 query-documents pairs upon which the relevance judgments were made. These relevance judgments are divided in three categories i.e. definitely relevant, possibly relevant, or not relevant. For experiments and evaluation, all the documents that are judged here as either possibly or definitely relevant were considered as relevant. Furthermore, only the documents to which the abstracts are available, were index and used for the retrieval task. Therefore, the experiments were conducted on the remaining set of 233,445 documents from the Ohsumed test collection. Also, to calculate the overall performance we considered only those queries, which had any relevant document(s) listed in the judgment file. Out of 106 queries in total, 97 of them were found to have relevant document(s) associated with it. This document collection is fairly large in terms document size but shorter in terms of document length as compare to the WebAP test collection. It does not include any annotation at passage level.

### 4.2 Assumptions and Experimental Parameters

For our experiments we used Solr-5.2.1 which is built on top of LUCENE[3]. Solr provided the functionality of removing the stop-words at indexing time. As shown in Fig. 1, we used that functionality to remove the stop words[4] from both collections. We have seen that the ranking after removing the stop-words is improved.

For different similarity measure functions, we used different parameters. For sum of passages (SF2) and inverse rank (SF4) function, we set the $k$ value to be equal to 5 and the results were normalized having received the final score. Similarly, we gave twice the boost to the passage level score as compared to the document level score while combining the results together i.e $\alpha = 1, \beta = 2$. Giving the higher boost to passage level gives better performance to the inverse ranking functions, whereas higher boost at document level improved results for Max passage and Sum of passage results.

### 4.3 Evaluation Measures

In IR, different evaluation measures are used to measure how well the system is performing to satisfy the user's need in returning the relevant documents to a given query. In our case, to measure the quality and performance of our approach, we used Mean Average Precision (MAP) and precision@k. MAP value is used to give an overall performance overview of the system and different similarity functions across both test collections. On the other hand, precision@k was helpful in illustrating the user's experience and the behavior of relevant documents returned in terms of their ranking frequency with the different threshold values. We evaluated the precision value for top 40 unique documents, both at passage level and at document level.

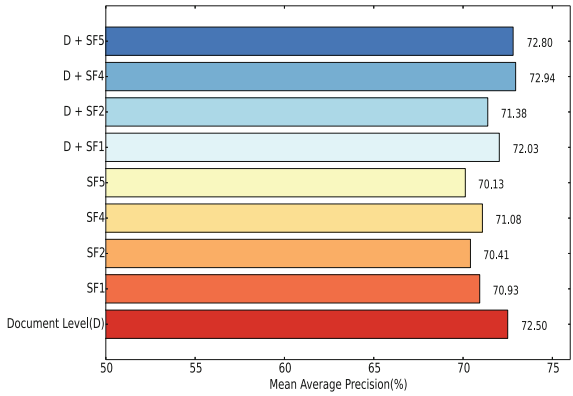
## 5 Results

In this section we present the experimental results to show the performance of the different similarity functions at passage level and document level for both the WebAP and Ohsumed datasets.

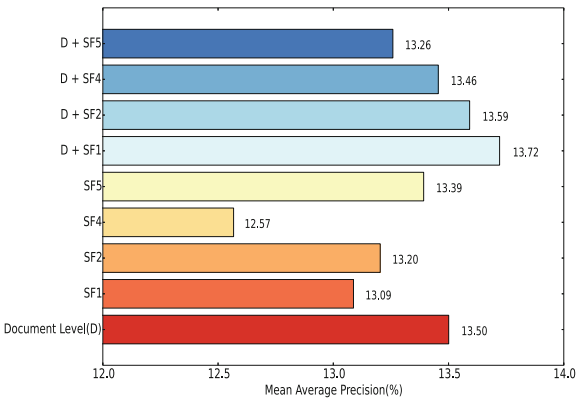
In Fig. 2(a) and (b), a bar chart is used to compare the document-level score with the different similarity functions of passage level scores for WebAP and Ohsumed test collections.

Using the WebAP collection, the results show that combining the document level score with passage level score (SF3), gives an improvement in performance. The best results were found when the document level score was combined with the inverse rank functions (SF4, SF5) of the passage level ranking. The results show that, considering the rank of the documents instead of the similarity score gives better performance when document ranking is combined with the passage

[3] http://lucene.apache.org/.
[4] http://www.ranks.nl/stopwords.




(a) WebAp Collection




(b) Ohsumed Collection

**Fig. 2.** Mean average precision for different similarity functions, extracted from [12].

level evidence. For the sum of passages (SF2) approach, only the top 5 (i.e. k = 5) results were considered in calculating the query similarity score.

In contrast to WebAP, for the Oushmed collection the combination of document score with the max passage score performed better than the combination of inverse passage rank with document score. However, for functions not including the document level similarity, inverse rank by alpha (SF5) performed better than the other passage level similarity functions and give approximately similar performance in comparison to document level. Furthermore, the sum of passages (SF2) performed better here than the Max passage (SF1) score. The best results were observed for $k = 2$. We have observed that the MAP values decrease as the $k$ value increases, hence max passage similarity function performs better than the sum of passages function for WebAP test collection. However, in Ohsumed SF2 performed better than SF1 for $k = \{2, 3, 4\}$.

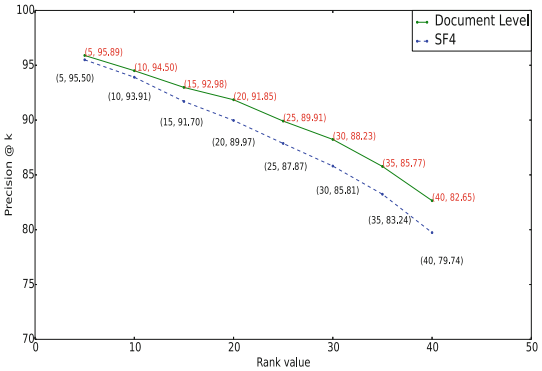
We also used precision@k as a different evaluation metric. The objective of this experiment was to check how well the documents are returned at the top $k$ ranks at the document and passage level, and to measure on average how many relevant documents are returned at the different $k$ values. Figure 3(a) and (b) illustrate the calculated precision values for WebAP test collection and the Ohsumed collection at document level as well as at passage level. At passage level we used SF4 and SF5 to measure the average precision for the WebAP and the Ohsumed, as when we considered it separately (without in conjunction with the document score), their performance was better than SF1 and SF2.

For the WebAP, the results show that the document level achieved better $p@k$ in comparison to SF4, and out of 40 documents, 33 of them are relevant in document level and 31 of them are relevant at the passage level when SF4 was used. On average, the precision value for document level and passage level was 90% and 86%. This indicates that the correct documents for all queries are clustered together or are closely related to each other and therefore, most of them are returned in top results, hence the high results.

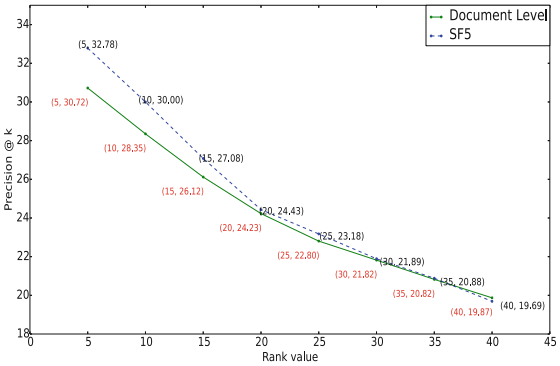
For the Ohsumed collection, SF5 clearly outperformed the document level results and gave marginally better precision from the start to top 20 results (p@20) compared to the document level. However, for the higher values i.e. $k > 20$ , the document level and SF5 gave almost the similar performance. Out of 40 documents approximately 9 are relevant in document retrieval and 10 of them are relevant in passage retrieval by using the inverse rank by alpha function (SF5). The overall performance for the Ohsumed collection is fairly low and this could be partially due to the large size of the test collection, small document length and the variation of relevant document information in relevance judgment file. On average, precision value for the document level and passage level was 24% and 25%.

Table 1 illustrates the mean average precision at top 5 (MAP@5) and at top 10 (MAP@10) for both test collections and as the results were discussed before, in the WebAP the combination of document level with passage level scores with different similarity functions give better results. The best results were obtained when the document score is combined with SF5. Whereas, for the Ohsumed, the functions that do not involve combining passage level and document level evidence gives better performance in both cases.

To get a better understanding on the statistical significance of the differences shown in the Table 1 for the test collections, we used the Student's t-test on paired samples for the top 50 MAP values with the difference of 5 (i.e. top 5, top 10, top 15, till top 50 etc). For the WebAP, we compared the document level results with the D+SF5 similarity function as it gave an overall better performance on the top results. The average MAP difference between both experiments was 0.18 with the standard deviation of 0.09 and the calculated p-value was 0.00024. Therefore, the performance shown by D+SF5 is statistically significant as compared to the normal document level results. Similarly, we performed the same t-test on the Ohsumed collection by comparing the document level results with D+SF4 due to its advantage over the performance on normal document




(a) WebAP Collection




(b) Ohsumed Collection

**Fig. 3.** Precision at K for different test collections, extracted from [12].

level results. For the Ohsumed, the average difference and standard deviation were 0.07 and 0.13 with the p-value of 0.069. Hence, for the Ohsumed, the results were not improved very significantly.

It is also seen that the value of $\alpha$ and $\beta$ effects the overall results when the document level is combined with the passage level evidence (SF3). For both collections, giving the higher boost to passage level i.e. $\alpha <= \beta$, gave a better performance for the inverse ranking functions, whereas a higher boost at document level i.e. $\alpha > \beta$ improves the results for SF1 and SF2. We chose $\alpha = 1$ and $\beta = 2$ for the results shown in this paper because it gives an overall better performance for all the passage level similarity functions when combined with the document score.

### 5.1 Further Analysis of SF2

In the previous section, we used $k = 2$ to report the results for SF2 using the WebAp and the Oshumed test collection. We have seen that by varying the

value of $k$, the average precision changes, which leads us to highlight the effect of changing the value of $k$ in SF2 against the test collections used in this paper. To understand, and to determine how well the addition of passages performed in terms of improving the document ranking, we illustrate the behavior of SF2 at different k values. We report the results for $k = 1, 2, 3, 4, 5$, because considering the average size of the number of passages per document in both test collections, bigger k values i.e >5 did not demonstrate any improvement in performance. Figure 4(a) and (b) shows how the Mean Average Precision changes for different k values in both the test collections.

For the WebAP collection we have seen that the precision decreases with increasing k values. However, for the Ohsumed collections the best value is obtained for k = 2. This could be due to the number of passages per document in both collections. The WebAp has documents with the bigger document length, having more than 10 passages per document on average.

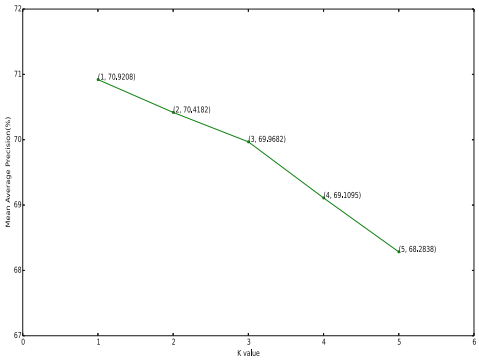
In the Ohsumed collection the document length is quite small with around 3–4 passages per document. Moreover, it is worth noting that by adding more passages together i.e., with the increase in the k value, we are losing the accuracy and adding more noise in the result set, which could be a cause of decrease in the MAP value. And as shown in Fig. 4(a), the higher values of k are giving lower precision in both collections that supports our argument regarding the decrease in efficacy and the increase in noise.

We have also performed the one sample T-test to check if the difference between the MAP at various k values is significant or not. For the WebAp collection, the p value <0.0001 and therefore, this difference is considered to be extremely statistically significant. Similarly, for the Ohsumed collection, the p values is also <0.0001, which makes the difference significant as well.

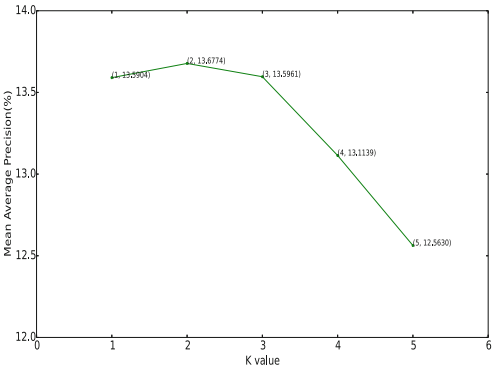
### 5.2 Query Difficulty

Oftentimes, information retrieval systems exhibit a substantial variance in accuracy across a set of queries. Systems may display a similar MAP, but quite a considerable variance in performance when considered in a query-by-query manner. A large body of work exists in predicting query performance; i.e. given a particular query, can one predict the expected MAP from a particular IR system? A range of techniques have been considered; these can be broadly categorised into two main categories: pre-retrieval and post-retrieval. Pre-retrieval techniques consider examining the query and looking at features of the query and the query term; these include linguistic approaches [23] and statistical approaches [24,29]. Post-retrieval, on the other hand, examines features of the returned answer and attempts to gauge the quality of the answer as a measure of the query difficulty. Researched approached include consider the distribution of similarity scores [30] and cohesion of the answer set [31].

In this work, we have explored a number of passage level approaches and demonstrated in some cases, a modest, yet significant improvement over the baseline adopting a classical document-level approach. However, it is not known if this improvement is due to a large number of slight improvements over a large range of queries or due to larger improvements over a small set of queries.




(a) WebAp Collection



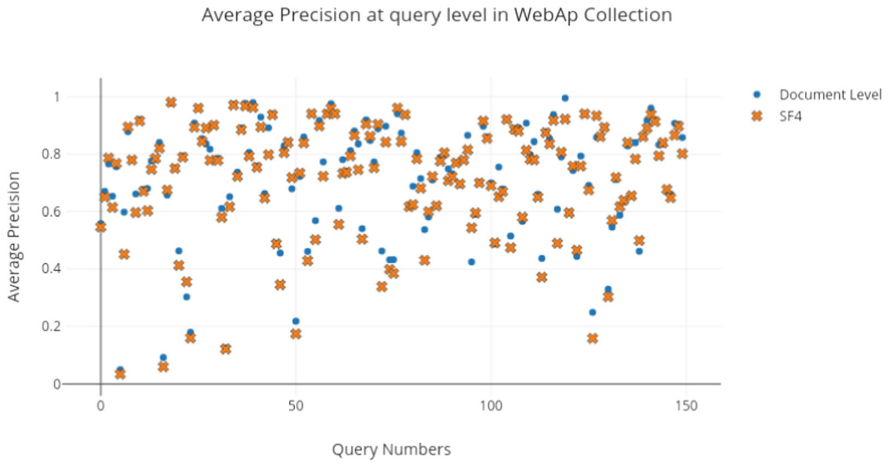

(b) Ohsumed Collection

**Fig. 4.** SF2 results for top 5 $K$ values.

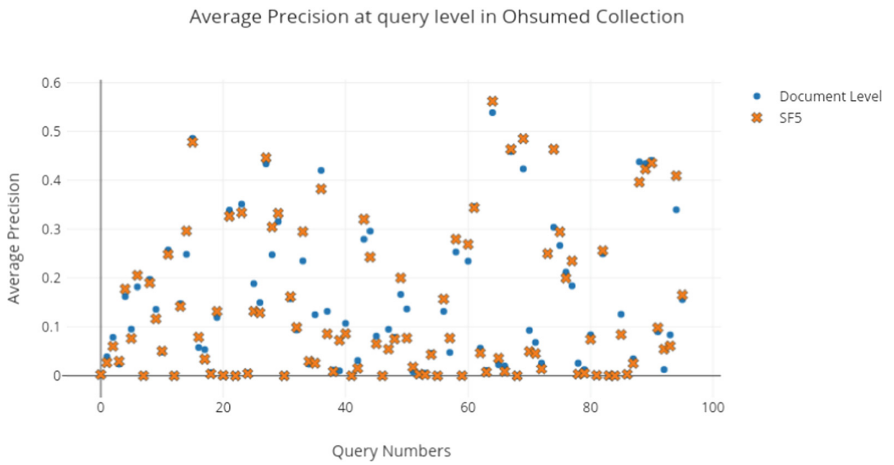
Moreover, it is worth exploring if the best performing passage level approaches actually damage the performance for certain queries.

In this section, we present a query by query overview of the performance of the baseline and the best passage level ranking function in both the collections. In the WebAP collection, the SF4 approach performed best and in the Ohsumed collection, the SF5 approach gave the better performance. We use these similarity functions to compare their impact on all the queries in their respective test collections. We identify the queries for which there is a substantial change in performance between the two and attempt to provide an explanation for this change.

Figure 5(a) and (b) illustrate the average precision across all queries. As shown in Fig. 2(a) and (b), we compare the baseline results with the similarity functions that was giving the better performance at passage level i.e. SF4 in the WebAp and the SF5 for the Ohsumed collection. In the WebAP collection, we saw that for the difficult queries (queries which performed worst), the document




(a) WebAp Collection




(b) Ohsumed Collection

**Fig. 5.** Average precision of each query for different test collections.

level was giving better performance than the SF4 approach. We also measured the query length of bottom 10 queries against document level and SF4 and didn't find any significant difference between them. On average the query length of the document level results and SF4 was 3.3 and 3.1 words per query. Moreover, for the WebAP collection, we have seen that the SF4 performed better because of the substantial improvements in a small subset of the queries (easy and difficult ones) and not due to large number of slight improvements over a set of queries. Of the 150 queries in total, in SF4 performed better than the baseline in 38 of them.

It appears that, for the passage level technique (SF4), the worst-performing queries are damaged slightly and the those that perform well are boosted. For the poorly performing queries, the IR system has difficulty distinguishing the relevant from non-relevant documents (similar term frequency distributions). In incorporating passage level evidence we are possibly including evidence from weakly related passages. Rather than improving performance, we are merely hampering performance by incorporating information that does not improve our ability to make a useful similarity estimate.

For the Ohsumed collection, due to very low values of average precision, we considered the bottom 20 queries in order to get the better understanding of how well the difficult queries are performing against the document level and SF5. For difficult queries, SF5 gave better performance against the document level results and the average number of words per query noted for SF5 and document level was 6.5 and 6.4 words. Though the difference between them is not significant but SF5 slightly boosted the results for worst-performing queries. Among the total 96 queries in the Ohsumed collection, SF5 gave better accuracy for the 42 queries (including 25 difficult queries) compare to the document level results. Hence, we can say that the overall performance is increased due the small improvements in the large set of difficult queries.

An overall better approach would be to attempt to identify, in advance, which queries are likely to be improved by the passage level augmentation. To this end, we attempt to identify differences between those queries that are benefited by the passage level and those whose performance is damaged.

## 6 Conclusions and Future Work

In this paper, the main focus of our work was to see how effectively the passage level evidence affected the document retrieval. We explored several similarity measures that can be used to improve the document ranking. Though we saw that the rank of a passage is an effective measure, however the passage level evidence on its own is not ample to improve the document ranking significantly for the selected test collections. In addition to that, we undertook the detailed analysis of SF2 to understand its behavior on different $k$ values. SF2 performed best when the value of k is smaller. For the WebAP collection, we notice that the precision decreases with increasing $k$ values. However, for the Ohsumed collection, the best value is obtained for $k = 2$. Moreover, we investigated the idea of query difficulty with regards to its impact on our rank-based passage functions. For the WebAp, we compared the baseline results with SF4, and SF5 was used in the Ohsumed collection due to its higher performance. Final results reveal that for the passage level technique, the difficult queries are damaged slightly and the those that perform well are boosted for the WebAp collection. However, for the Ohsumed collection, SF5 promoted the performance of the worst-performing queries. Given the evidence that passage level evidence can improve the performance and given the results to show that the level of improvement often depends on the query difficulty, future work will explore other passage level evidence and

also query difficulty estimation approaches to attempt to develop a more nuanced approached to ranking using passage level evidence in scenarios where the difficulty of the query can be estimated.

**Acknowledgements.** This work is supported by the Irish Research Council Employment Based Programme.

## References

1. Robertson, S., Zaragoza, H., et al.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends ® Inf. Retr. **3**, 333–389 (2009)
2. Roberts, I., Gaizauskas, R.: Evaluating passage retrieval approaches for question answering. In: McDonald, S., Tait, J. (eds.) ECIR 2004. LNCS, vol. 2997, pp. 72–84. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24752-4_6
3. Sarwar, G., O'Riordan, C., Newell, J.: Passage level evidence for effective document level retrieval. In: Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, pp. 83–90 (2017)
4. Hersh, W., Buckley, C., Leone, T., Hickam, D.: OHSUMED: an interactive retrieval evaluation and new large test collection for research. In: Croft, B.W., van Rijsbergen, C.J. (eds.) SIGIR 1994, pp. 192–201. Springer, London (1994). https://doi.org/10.1007/978-1-4471-2099-5_20
5. Callan, J.P.: Passage-level evidence in document retrieval. In: Croft, B.W., van Rijsbergen, C.J. (eds.) SIGIR 1994, pp. 302–310. Springer, London (1994). https://doi.org/10.1007/978-1-4471-2099-5_31
6. Hearst, M.A.: Texttiling: segmenting text into multi-paragraph subtopic passages. Comput. Linguist. **23**, 33–64 (1997)
7. Bendersky, M., Kurland, O.: Utilizing passage-based language models for document retrieval. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 162–174. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78646-7_17
8. Kaszkiel, M., Zobel, J.: Effective ranking with arbitrary passages. J. Am. Soc. Inf. Sci. Technol. **52**, 344–364 (2001)
9. Clarke, C.L., Cormack, G.V., Lynam, T.R., Terra, E.L.: Question answering by passage selection. In: Strzalkowski, T., Harabagiu, S.M. (eds.) Advances in Open Domain Question Answering, pp. 259–283. Springer, Dordrecht (2008). https://doi.org/10.1007/978-1-4020-4746-6_8
10. Liu, X., Croft, W.B.: Passage retrieval based on language models. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management, pp. 375–382. ACM (2002)
11. Jong, M.H., Ri, C.H., Choe, H.C., Hwang, C.J.: A method of passage-based document retrieval in question answering system. arXiv preprint arXiv:1512.05437 (2015)
12. Sarwar, G., O'Riordan, C., Newell, J.: Passage level evidence for effective document level retrieval (2017)
13. Buckley, C., Salton, G., Allan, J., Singhal, A.: Automatic query expansion using smart: TREC 3. NIST Special Publication SP, p. 69 (1995)
14. Hearst, M.A., Plaunt, C.: Subtopic structuring for full-length document access. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 59–68. ACM (1993)

15. Salton, G., Allan, J., Buckley, C.: Approaches to passage retrieval in full text information systems. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 49–58. ACM (1993)
16. Lavrenko, V., Croft, W.B.: Relevance based language models. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and development in Information Retrieval, pp. 120–127. ACM (2001)
17. Krikon, E., Kurland, O., Bendersky, M.: Utilizing inter-passage and inter-document similarities for reranking search results. ACM Trans. Inf.Syst. (TOIS) **29**, 3 (2010)
18. Bendersky, M., Kurland, O.: Re-ranking search results using document-passage graphs. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 853–854. ACM (2008)
19. Ai, Q., O'Connor, B., Croft, W.B.: A neural passage model for ad-hoc document retrieval. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) ECIR 2018. LNCS, vol. 10772, pp. 537–543. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76941-7_41
20. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 275–281. ACM (1998)
21. Galkó, F., Eickhoff, C.: Biomedical question answering via weighted neural network passage retrieval. arXiv preprint arXiv:1801.02832 (2018)
22. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
23. Mothe, J., Tanguy, L.: Linguistic features to predict query difficulty
24. He, B., Ounis, I.: Inferring query performance using pre-retrieval predictors. In: Apostolico, A., Melucci, M. (eds.) SPIRE 2004. LNCS, vol. 3246, pp. 43–54. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30213-1_5
25. Lashkari, A.H., Mahdavi, F., Ghomi, V.: A boolean model in information retrieval for search engines. In: International Conference on Information Management and Engineering, ICIME 2009, pp. 385–389. IEEE (2009)
26. Keikha, M., Park, J.H., Croft, W.B., Sanderson, M.: Retrieving passages and finding answers. In: Proceedings of the 2014 Australasian Document Computing Symposium, p. 81. ACM (2014)
27. Chen, R.C., Spina, D., Croft, W.B., Sanderson, M., Scholer, F.: Harnessing semantics for answer sentence retrieval. In: Proceedings of the Eighth Workshop on Exploiting Semantic Annotations in Information Retrieval, pp. 21–27. ACM (2015)
28. Yang, L., et al.: Beyond factoid QA: effective methods for non-factoid answer sentence retrieval. In: Ferro, N., et al. (eds.) ECIR 2016. LNCS, vol. 9626, pp. 115–128. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30671-1_9
29. He, J., Larson, M., de Rijke, M.: Using coherence-based measures to predict query difficulty. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 689–694. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78646-7_80
30. Cummins, R., Jose, J., O'Riordan, C.: Improved query performance prediction using standard deviation. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, pp. 1089–1090. ACM, New York (2011)
31. Vinay, V., Cox, I.J., Milic-Frayling, N., Wood, K.: On ranking the effectiveness of searches. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2006, pp. 398–404. ACM, New York (2006)