

# Investigating Passage-level Relevance and Its Role in Document-level Relevance Judgment

Zhijing Wu, Jiaxin Mao, Yiqun Liu\*, Min Zhang, Shaoping Ma  
Department of Computer Science and Technology, Institute for Artificial Intelligence,  
Beijing National Research Center for Information Science and Technology,  
Tsinghua University, Beijing 100084, China  
yiqunliu@tsinghua.edu.cn

## ABSTRACT

The understanding of the process of relevance judgment helps to inspire the design of retrieval models. Traditional retrieval models usually estimate relevance based on document-level signals. Recent works consider a more fine-grain, passage-level relevance information, which can further enhance retrieval performance. However, it lacks a detailed analysis of how passage-level relevance signals determine or influence the relevance judgment of the whole document. To investigate the role of passage-level relevance in the document-level relevance judgment, we construct an *ad-hoc* retrieval dataset with both passage-level and document-level relevance labels. A thorough analysis reveals that: 1) there is a strong correlation between the document-level relevance and the fractions of irrelevant passages to highly relevant passages; 2) the position, length and query similarity of passages play different roles in the determination of document-level relevance; 3) The sequential passage-level relevance within a document is a potential indicator for the document-level relevance. Based on the relationship between passage-level and document-level relevance, we also show that utilizing passage-level relevance signals can improve existing document ranking models. This study helps us better understand how users perceive relevance for a document and inspire the designing of novel ranking models leveraging fine-grain, passage-level relevance signals.

## KEYWORDS

Relevance judgment; passage-level relevance aggregation; relevance model

### ACM Reference Format:

Zhijing Wu, Jiaxin Mao, Yiqun Liu, Min Zhang, Shaoping Ma. 2019. Investigating Passage-level Relevance and Its Role in Document-level Relevance Judgment. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3331184.3331233>

\*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

SIGIR '19, July 21–25, 2019, Paris, France

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6172-9/19/07...\$15.00

<https://doi.org/10.1145/3331184.3331233>

## 1 INTRODUCTION

Understanding the process of relevance judgment of human for a specific query-document pair is essential, which can inspire a better design of relevance-based ranking models. Traditional document ranking models usually estimate relevance based on document-level signals [1, 28]. To analyse further, a document has a hierarchical internal structure. It is composed of multiple passages, which can be separated from the whole document based on textual discourse unit (discourse passage), subject or content of the text (semantic passage), or a certain number of words (window passage) [2]. A document could be partially relevant to a query as long as it provides a certain amount of useful information like some relevant passages for users' information needs. Several works attempted to rank documents by using passage-level relevance information and all of them illustrated that incorporating passage-level relevance signals can enhance the performance in document ranking [4, 16]. Therefore, investigating passage-level relevance and document-level relevance sheds light on the understanding of the process of relevance judgment and may benefit the design of ranking models.

Numerous existing works have tried to study how to estimate document relevance with passage-level relevance. Liu and Croft [22] took the highest passage-level relevance score of all passages as the document-level relevance score. Kong et al. [18] compared the retrieval effectiveness of three different decision principles: aggregate relevance (AR) principle, disjunctive relevance decision (DRD) principle, and conjunctive relevance decision (CRD) principle. AR assumes the more relevant passages in a document, the higher the relevance score of the document. In particularly, DRD and CRD principles are two extreme cases of AR. With DRD/CRD, if a passage in a document is relevant/irrelevant, then the entire document is relevant/irrelevant. Kong et al. [18] found that the generalized mean aggregation operator derived from the AR principle is the best choice for estimating the document-level relevance. Wilkinson [34] split the document into passages based on textual discourse units (i.e., sections) and showed that considering the type of each section also promotes the ranking performance. Recently, Fan et al. [4] proposed a neural model to learn relevance signals at different granularities (i.e., passage-level and document-level). Inel et al. [14] showed that aggregating passage-level relevance can boost the accuracy of relevance estimation.

However, judging the relevance of the whole document can be an intricate process. Existing studies on aggregation methods of the passage-level relevance are heuristic and ignore some potential factors that affect relevance judgment. For instance, Li et al. [20] conducted an eye-tracking study to demonstrate that the passages at the beginning of the document attract more attention and have

more substantial influence in determining the overall document-level relevance. There is a lack of detailed analysis of how the relevance judgment of the whole document is determined or influenced by the fine-grain, passage-level relevance signals. Therefore, we systematically investigate how the local passage-level relevance affects the global document-level relevance. Specifically, we try to address the following research questions:

- **RQ1:** What is the relationship between document-level relevance and the relevance of passages composed of the document?
- **RQ2:** Can we promote the performance of document ranking with the help of this relationship?

To address these research questions, we build a dataset with 70 queries sampled from the search logs of a commercial Web search engine and 1,050 related documents (15 documents per query) retrieved from a Chinese news corpus<sup>1</sup>. For the 1,050 query-document pairs and 11,512 query-passage pairs within the documents, we collect four-grade relevance judgments from well-trained workers. From the dataset, we first analyse the relationship between passage-level relevance and document-level relevance to address RQ1. Based on the findings in RQ1, we then incorporate the passage-level relevance signals into existing document ranking models (RQ2). Experimental results show that there is a significant improvement in terms of document ranking performance. To summarize, the main contributions are as follows:

- We construct an annotated dataset<sup>2</sup> for four-grade relevance judgments consisting of 11,512 passages and 1,050 documents.
- We provide a thorough analysis of how people perceive the document-level relevance from passage-level relevance, which sheds light on the understanding of how people make global relevance judgments based on local relevance.
- We show that deploying passage-level relevance signals improves the existing document ranking models.

The rest of this paper is outlined as follows. In Section 2, we review related work. In Section 3, we describe the dataset used in our study. In Section 4, we analyse the relationship between passage-level and document-level relevance to address RQ1. In Section 5, we build models for document ranking.

## 2 RELATED WORK

Making relevance judgment for a certain query-document pair is an intricate process. Wu et al. [35] formulate this process as firstly making relevance decision at specified locations in the document, then incorporating the relevance at each location as the final global document relevance score. We consider the passage-level relevance as the local relevance in this paper. Therefore, we briefly review the related work on 1) passage-level relevance and 2) relevance estimation with passage-level relevance.

**Passage-level Relevance.** Passage-level relevance is one type of local relevance within the document. Previous works have introduced three types of local relevance: 1) query-centric context relevance [25, 35, 36]; 2) field-level relevance [26, 27, 39]; 3) passage-level relevance [2, 4, 11, 16, 22, 35]. Based on the assumption proposed by Wu et al. [35], the query-centric context considers that

the relevant information must locate around the query terms inside the document. Therefore the query-centric context is defined as a context with a query term at the center position. The field-level relevance considers that several fields of the document (e.g., headlines, main text, anchor text) are of different importance in the document-level relevance judgment. Different from the query-centric context relevance and field-level relevance, passage-level relevance considers the relevance scores of passages within the document, which can be grouped into semantic, window, and discourse passages [2]. Semantic passages are derived from documents by algorithms such as TextTiling [10] based on the subject or content of the text. Window passages consist of a fixed number of words or bytes [17, 37], which may not take logical structure of the document into account. Discourse passages are based on textual discourse units such as sentences, paragraphs, and sections [11, 29]. In this paper, we construct the dataset from a News corpus, where the documents are well-organized according to its logical structure. We can split documents into passages based on the textual discourse units. Therefore, we mainly focus on the paragraph passages here.

Several works have investigated the relevance judgment at passage level. White et al. [33] and Callan [2] propose that it is natural to consider the fine-grained relevance such as passage-level relevance with the increase of documents' length. It avoids the difficulties of comparing documents of different length and is proved to be robust and effective in document retrieval [16]. Trotman et al. [31] and McDonnell et al. [23] require assessors to annotate relevant passages for supporting the judgment of document relevance. Recently, Inel et al. [14] use a binary relevance scale to annotate each paragraph passage of a document. In this paper, we collect four-grade passage-level relevance judgments for query-passage pairs.

**Relevance Estimation with Passage-level Relevance.** A large number of relevance estimation methods for document ranking have been proposed in the past few decades. Traditional methods usually consider the document as a whole and estimate its relevance based on document-level signals, such as BM25 model [28], pointwise, pairwise, and listwise learning to rank models [1, 3, 6, 8, 15, 38] and some deep learning models [9, 12, 13, 30]. Several works try to rank documents using passage-level relevance information. They demonstrate that incorporating passage-level relevance signals enhances the performance in document ranking [4, 16]. Liu and Croft [22] choose the highest passage-level relevance score of all passages as the document-level relevance score. Wang and Si [32] combined the document retrieval results with passage retrieval results using heuristic functions. Kong et al. [18] compare the retrieval effectiveness of three different decision principles: aggregate relevance (AR) principle, disjunctive relevance decision (DRD) principle, and conjunctive relevance decision (CRD) principle. They find that the generalized mean aggregation operator derived from the AR principle is the best choice. Recently, Fan et al. [4] propose a neural model that utilizes both the passage-level and document-level matching signals for document ranking and show that this model significantly outperforms existing ranking models.

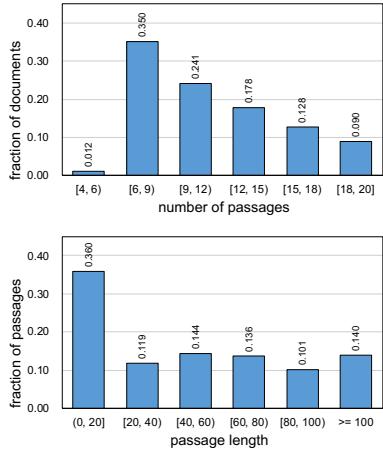
However, there is still a lack of thorough analysis and explanation of the relationship between document-level relevance and the relevance of passages composed of the document.

<sup>1</sup><http://thuctc.thunlp.org/>

<sup>2</sup>The data is now available at <http://www.thuir.cn/group/~YQLiu/>

**Table 1: Search query examples (translated from Chinese).**

| Query                             | Domain        | Query Type   | Search Background  |
|-----------------------------------|---------------|--------------|--|
| IELTS speaking test standard      | Education     | Factual      | You are preparing for the IELTS test recently and want to know the standard of IELTS speaking. |
| ONE PIECE                         | Entertainment | Factual      | You want to know some activity information of the mobile game ONE PIECE.                       |
| Reasons for the rise of oil price | Finance       | Intellectual | Recently the oil price has risen. You want to know the possible reasons behind it.             |
| Tips for kitchen decoration       | Lifestyle     | Intellectual | You are preparing to decorate the house and want to know some tips for kitchen decoration.     |



**Figure 1: The distributions of the number of passages within a document and the number of words within a passage.**

### 3 DATASET

In this section, we describe how we construct the dataset that we use throughout this paper and its statistics. The procedure of data collection consists of two steps. The first step is to construct the set of query-document pairs. Secondly, we hire external well-trained assessors to make relevance judgments for the documents and passages within the documents. We will release our dataset including queries, documents, and the relevance annotations at the document and passage levels after the review process.

#### 3.1 Constructing Dataset

To construct the query-document pairs for our experiment, we use THUCNews<sup>3</sup>, a Chinese news dataset, as our corpus and select queries from a 10-day query logs of a popular commercial search engine in China. The THUCNews corpus is based on the Web pages data of Sina News RSS subscription channel<sup>4</sup>. It includes 740 thousand well-organized and of high quality news documents with extracted full-text content. These documents cover 14 domains (e.g., education, entertainment, and finance). In this paper, we directly use one paragraph within the document as a passage (i.e., the paragraph passages introduced by [2]).

<sup>3</sup><http://thuctc.thunlp.org/>

<sup>4</sup><http://rss.sina.com.cn/>

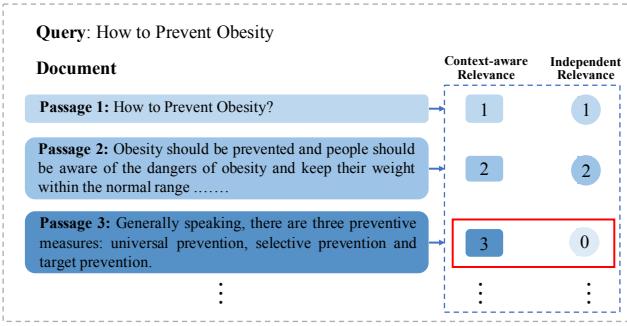
**Table 2: The statistics of the dataset.** The #Q, #D, #P, #P/D, and #W/P mean the number of queries, documents, passages, the average number of passages within a document, and the average number of words within a passage respectively.

| Type          | #Q | #D    | #P     | #P/D | #W/P |
|---------------|----|-------|--------|------|------|
| Education     | 16 | 240   | 2,697  | 11.2 | 50.6 |
| Entertainment | 30 | 450   | 4,739  | 10.5 | 53.6 |
| Finance       | 16 | 240   | 2,792  | 11.6 | 49.7 |
| Lifestyle     | 4  | 60    | 716    | 11.9 | 51.1 |
| Technology    | 4  | 60    | 568    | 9.5  | 43.6 |
| Factual       | 44 | 660   | 6,994  | 10.6 | 51.5 |
| Intellectual  | 26 | 390   | 4,518  | 11.6 | 51.0 |
| All           | 70 | 1,050 | 11,512 | 11.0 | 51.3 |

To select news-related queries, we sample the search sessions in the query logs where users have clicked on at least one result from the Sina news website<sup>5</sup> and reserve the corresponding queries. Considering that queries of low frequency are not usually used in users' daily life and search engines work well enough on the high-frequency queries, we manually choose 70 intermediate-frequency queries as our query set from these queries. These 70 queries cover five domains: education, entertainment, finance, lifestyle, and technology. For each query, we create a description of search background to make the query intent more clear and unambiguous. Considering that the process of relevance judgment is affected by the search task types [20], we group these queries into factual and intellectual categories according to the criteria introduced by Li and Belkin [21]. A factual query is submitted to the search engine for locating facts, data, or other similar information items, while an intellectual query is submitted for seeking new ideas or findings. There are 44 factual queries and 26 intellectual queries respectively. Table 1 shows four query examples and their corresponding background descriptions, which we translate from Chinese into English.

We describe the document sampling method in this study. The initial candidate document set consists of all the documents in THUCNews of the same domain with the query. We filter out the documents where the number of paragraphs is less than 4 or more than 20, which are too short or too long in the corpus. We calculate the BM25 score for each query-document pair, where the

<sup>5</sup>[sina.com.cn](http://sina.com.cn)



**Figure 2: Examples of query-passage pairs with context-aware and independent passage-level relevance annotations.**

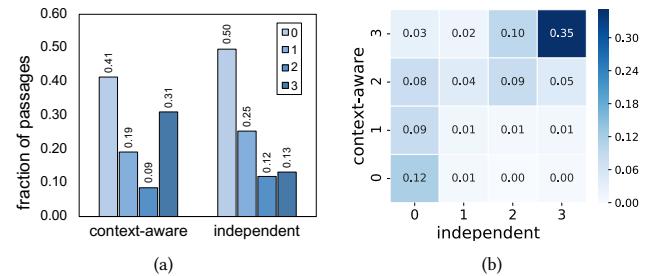
inverse document frequency values (IDF) of query words are estimated on the whole THUCNews corpus. Then, the documents are ranked according to their BM25 scores. Since the relevance annotation for each query-passage pair within a document is rather time-consuming and expensive, we randomly sample 15 documents from the top 20 results as the document set for the query. Considering that the relevance scores of the top 15 documents are high and there may be just a small number of irrelevant documents, we have not directly used the top 15 documents.

Finally, we obtain a dataset of 1,050 documents and 11,512 passages for the following experiment. The statistics of the dataset is shown in Table 2. There are 11.0 passages within a document and 51.3 words within a passage on average. The documents in the technology domain is shorter than those of the other domains. The distributions of the document length (i.e., the number of passages) and the passage length (i.e., the number of words) are shown in Figure 1. About 35% documents contain fewer than 10 passages and there are 36% passages that contain fewer than 20 words, which are usually the news leads or sub-headings of the documents.

### 3.2 Relevance Annotation

With the dataset introduced above, we collect relevance judgments for both the query-document and query-passage pairs from well-trained assessors. The assessors are hired by a crowdsourcing platform and are familiar with the relevance annotation task. Two groups of assessors are employed to make relevance judgments for the documents and passages respectively. They are required to examine the query, the description of search background, and the document/passage, then make a four-grade relevance judgment for the query-document or query-passage pair. Before the formal annotation task, they need to do some training annotation to make sure that they have correctly understood the annotation rules. Each document/passage is annotated by three assessors. The relevance scales and instructions are as follows:

- **(0) Irrelevant.** The content of the document/passage is not related to the subject of the query at all.
- **(1) Marginally relevant.** The topic of the query is mentioned, but only in passing and the objects still cannot satisfy the information needs.



**Figure 3: The (a) marginal distributions and (b) joint distribution of context-aware and independent passage-level relevance judgments in the pilot annotation. The values from 0 to 3 mean respectively irrelevant, marginally relevant, fairly relevant, and highly relevant.**

- **(2) Fairly relevant.** The topic of query is discussed. Based on the object, the information needs can be fairly satisfied.
- **(3) Highly relevant.** The topic is the main theme of the document/passage. Based on the object, the information needs can be totally satisfied.

In the document-level relevance annotations, all 1,050 query-document pairs are shown to the assessors in random order. Compared to the documents, the passages are much shorter and provide less information. In the context of a document, the passages usually contain metonymies or pronouns, which may be ambiguous and confusing without the context information. However, Inel et al. [14] show that with the context information, assessors tend to make the same relevance judgment for the current passage and the previous passage, which may lead to biased relevance judgments. To investigate whether the context information should be provided during the annotation process, we conduct a pilot annotation of passage-level relevance on a small dataset, which contains 14 queries and 56 corresponding news documents (4 documents per query). We test the following two settings in the pilot passage-level relevance annotation:

- **Context-aware passage-level relevance annotation (CRA):** show the whole document to the assessor and ask her to make relevance judgments for all passages within it.
- **Independent passage-level relevance annotation (IRA):** each time show only one passage to the assessor and ask her to make a relevance judgment for it.

There are 419 passages in the small dataset. Each passage is also annotated by three assessors. The Fleiss'  $\kappa$  [5] among the assessors for CRA and IRA are 0.702 and 0.523 respectively. The higher  $\kappa$  in the CRA setting suggests that the assessors are easier to reach an agreement on relevance judgment when the context information is given. We use the median relevance scores of three assessors as the relevance label of the passage. Figure 3 shows the marginal distributions and joint distribution in the CRA and IRA settings. We find that in the CRA setting, the fraction of highly relevant passages to all passages is higher than that under the IRA setting. One example of query-passage pair which has different relevance annotations in the CRA and IRA settings is shown in Figure 2. We

**Table 3:** The distributions of four-grade document-level relevance ( $r_d$ ) and passage-level relevance ( $r_p$ ). The Avg. #P means the average number of passages within the documents. The Avg. #W means the average number of words within the passages.

| Document  |          |         | Passage   |          |         |
|-----------|----------|---------|-----------|----------|---------|
| Type      | Fraction | Avg. #P | Type      | Fraction | Avg. #W |
| $r_d = 0$ | 0.472    | 10.9    | $r_p = 0$ | 0.724    | 44.3    |
| $r_d = 1$ | 0.161    | 11.0    | $r_p = 1$ | 0.137    | 61.9    |
| $r_d = 2$ | 0.115    | 11.2    | $r_p = 2$ | 0.044    | 74.3    |
| $r_d = 3$ | 0.251    | 11.0    | $r_p = 3$ | 0.096    | 78.3    |
| All       | 1        | 11.0    | All       | 1        | 51.3    |

find that without the context information, assessors can not decide what “preventive measures” refers to. Consequently, they make irrelevant annotations on this query-passage pair, which is less accurate than annotations with context information. We use the CRA setting to make the passage-level relevance annotation for our main dataset.

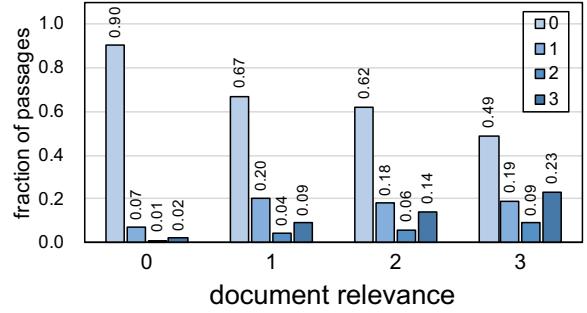
The Fleiss’  $\kappa$  of document-level relevance annotations and passage-level relevance annotations are 0.703 and 0.651, showing a substantial agreement among the assessors according to Landis and Koch [19]. We use the median relevance scores of three assessors as the relevance labels for the document and passage. Table 3 shows the distributions of document-level relevance and passage-level relevance. It shows that in our dataset, about 25% documents are highly relevant to the queries. The fraction of highly relevant passages is about 10%, which is lower than the fraction of highly relevant documents. The difference between the average lengths of documents with different relevance scores are negligible. However, at the passage-level, the irrelevant passages are shorter than the other passages on average.

## 4 DOCUMENT-LEVEL AND PASSAGE-LEVEL RELEVANCE JUDGMENTS

In this section, we investigate how people perceive document-level relevance from passage-level relevance to answer **RQ1**. The analysis includes the relationships between the document-level relevance and the distribution, weighted aggregation, as well as sequences of the passage-level relevance. We also analyse the effect of query types (i.e., factual and intellectual) on these relationships.

### 4.1 Passage-level Relevance Distribution

We first look into the distributions of passage-level relevance among documents with different relevance scores. Figure 4 shows the fractions of passages at different relevance levels in (0) irrelevant, (1) marginally relevant, (2) fairly relevant, and (3) highly relevant documents respectively. We find that the fraction of irrelevant passages within the document declines sharply as the document-level relevance increases. It indicates that with more irrelevant passages, the document is more likely to be judged as irrelevant by people. Within the highly relevant documents, there are still 49% irrelevant passages on average, which is inconsistent with the assumption of CRD principle proposed by Kong et al. [18] (If a



**Figure 4:** The joint distribution of document-level and passage-level relevance.

passage within the document is irrelevant, the entire document is irrelevant). The fraction of highly relevant passages within the document also increases as the document-level relevance increases. It agrees with the AP principle which assumes that if there are more relevant passages in a document, the relevance score of the document is higher. There are a few highly relevant passages (2%) within the irrelevant documents. People may not notice the highly relevant passages when making relevance judgments on documents that contain too many irrelevance passages. 23% passages within the highly relevant documents are highly relevant to the query, which can mostly satisfy the information needs individually. It shows that sometimes people can be satisfied with only reading a few passages rather than the entire document.

Based on these findings, we can use the fractions of irrelevant, marginally relevant, fairly relevant, and highly relevant passages to estimate the document-level relevance. To evaluate the performance of these estimation methods, we try to rerank the documents within a query. We calculate the Spearman’s correlation coefficient between the annotated document-level relevance and the fractions of passages at different relevance levels for each query, then we report the average coefficient of all queries in Table 4. As the top four rows of Table 4 show, there is a negative correlation between the fraction of irrelevant passages and the document-level relevance. We rerank the documents according to the fraction of irrelevant passages in ascending order and report the performance of nDCG@{5, 10, 15}. There is a positive correlation between the document-level relevance and the fraction of marginally/fairly/highly relevant passages. We also rerank the documents of one query according to the fraction of marginally/fairly/highly relevant passages in descending order. We find that these three methods perform worse than the method using the fraction of irrelevant passages. Therefore, in this experiment, we find that the fractions of irrelevant and highly relevant passages have the greatest impact on the document-level relevance.

### 4.2 Passage-level Relevance Aggregation

The document-level relevance is incorporated by pieces of local relevance [35]. In this study, we consider passage-level relevance as the local relevance. Previous works try to use the minimum, maximum, or simple aggregation of passage-level relevance as the document-level relevance score [18, 22]. Besides these methods,

**Table 4: Performance of distribution, aggregation, and sequence methods on the document-level relevance estimation.**

| Category     | Method                                       | Spearman     | nDCG@5       | nDCG@10      | nDCG@15      | AUC          |
|--------------|--|--------------|--------------|--------------|--------------|--------------|
| Distribution | the fraction of irrelevant passages          | -0.568       | <b>0.776</b> | <b>0.836</b> | <b>0.880</b> | <b>0.817</b> |
|              | the fraction of marginally relevant passages | 0.326        | 0.643        | 0.741        | 0.813        | 0.652        |
|              | the fraction of fairly relevant passages     | 0.337        | 0.656        | 0.708        | 0.816        | 0.655        |
|              | the fraction of highly relevant passages     | 0.515        | 0.761        | 0.826        | 0.879        | 0.759        |
| Aggregation  | minimum                                      | 0.041        | 0.461        | 0.583        | 0.719        | 0.522        |
|              | maximum                                      | 0.586        | 0.765        | 0.826        | 0.883        | 0.817        |
|              | median                                       | 0.422        | 0.691        | 0.737        | 0.846        | 0.706        |
|              | mean   | 0.622        | 0.832        | 0.871        | 0.911        | 0.841        |
|              | position decay                               | <b>0.624</b> | 0.835        | <b>0.877</b> | <b>0.913</b> | <b>0.845</b> |
|              | passage length                               | 0.604        | 0.832        | 0.867        | <b>0.913</b> | 0.836        |
|              | length with position decay                   | 0.614        | 0.827        | 0.866        | 0.910        | 0.841        |
|              | exact match                                  | 0.566        | 0.814        | 0.840        | 0.901        | 0.811        |
|              | query similarity                             | 0.620        | <b>0.836</b> | 0.870        | 0.909        | 0.841        |
| Sequence     | sub-sequence                                 | 0.594        | 0.787        | 0.846        | 0.889        | 0.835        |

we try more weighted aggregation methods in this section. Considering that some factors such as the position, length, and query similarity of passages may affect the importance of the passage in the document-level relevance judgment, we calculate different weights for the aggregation of passage-level relevance. A document  $d$  is represented as a set of passages  $d = \{p_1, p_2, p_3, \dots, p_n\}$ , where  $n$  denotes the number of passages in the document. Then, the estimated document-level relevance is calculated by the weighted aggregation of passage-level relevance:

$$\tilde{r}_d = \frac{\sum_{i=1}^n weight_i \times r_{p_i}}{\sum_{i=1}^n weight_i} \quad (1)$$

where  $r_{p_i}$  denotes the relevance score (0: irrelevant, 1: marginally relevant, 2: fairly relevant, 3: highly relevant) of the  $i$ -th passage,  $weight_i$  denotes the weight of the  $i$ -th passage. The estimated document-level relevance  $\tilde{r}_d$  is divided by the sum of  $weight_i$  to be normalized to [0, 3]. We try five aggregation methods using different factors as the weight: position, passage length, exact matching and query similarity. We describe how we calculate different weights as follows:

$$\tilde{r}_d = \frac{\sum_{i=1}^n r_{p_i}/i}{\sum_{i=1}^n 1/i} \quad (2)$$

**Position decay.** Li et al. [20] find that the passages at the beginning of the document attract more attention through an eye-tracking study. We assume that passages in the top positions have a greater influence in determining the overall document-level relevance. Therefore, we use the reciprocal of position as the weight of a passage, which is formulated as follows:

the  $i$ -th passage.

$$\tilde{r}_d = \frac{\sum_{i=1}^n |p_i| \times r_{p_i}}{\sum_{i=1}^n |p_i|} \quad (3)$$

**Length with position decay.** We further take the position and length of a passage into consideration at the same time and use the length with position decay as the weight.

$$\tilde{r}_d = \frac{\sum_{i=1}^n |p_i|/i \times r_{p_i}}{\sum_{i=1}^n |p_i|/i} \quad (4)$$

**Exact match.** Wu et al. [35] propose that the context with a query word at the center position provides strong signals for the document-level relevance judgment. Based on this assumption, we assume that the exact matching signal of a passage may affect its role in document-level relevance judgment. People may pay more attention to the passages with query words [20]. The aggregation of passage-level relevance is calculated as Equation 5, where  $O_{q,p}$  denotes the number of overlapping words both appearing in the query and the passage.

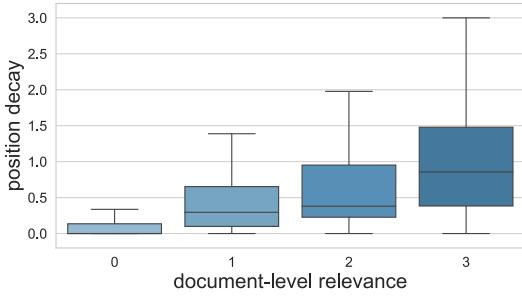
$$\tilde{r}_d = \frac{\sum_{i=1}^n O_{q,p} \times r_{p_i}}{\sum_{i=1}^n O_{q,p}} \quad (5)$$

**Query similarity.** As an extension of the *exact match* weighting method, we use the query similarity between the query and passage as the aggregation weight. We train the word embeddings on THUCNews corpus using word2vec [24] to map the semantic meaning of a word to a numerical representation. We use the average cosine similarity between the query and words within  $i$ -th passage as the weight, where the query ( $V_q$ ) is represented by the TF-IDF based weighted summation of query words.

$$V_q = \frac{\sum_{w \in q} IDF_w \times V_w}{\sum_{w \in q} IDF_w}$$

$$weight_i = \frac{\sum_{w \in p_i} \text{cosine\_similarity}(V_w, V_q)}{|p_i|}$$

$$\tilde{r}_d = \frac{\sum_{i=1}^n weight_i \times r_{p_i}}{\sum_{i=1}^n weight_i} \quad (6)$$



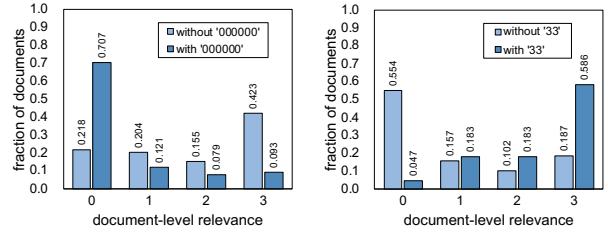
**Figure 5: The distributions of document-level relevance estimated by the *position decay* method, which belongs to the weighted aggregation category.**

In Table 4, we report the Spearman’s correlation coefficient between annotated and estimated document-level relevance. We rerank the documents in one query according to the estimated document-level relevance and report the nDCG at position 5, 10, and 15. Furthermore, we map the annotated document-level relevance labels to two-grade scores (0 and 1: irrelevant; 2 and 3: relevant) and scale the estimated document-level relevance scores to [0, 1]. The performances of different methods on AUC are reported in Table 4. We find significantly positive correlations between the document-level relevance and weighted aggregation results except for the minimum of passage-level relevance. Among the methods of *minimum*, *maximum*, *median*, and *mean*, *mean* method performs better than the others on document ranking, followed by *maximum*. We can see that *maximum* can achieve better performance on nDCG@15 than all the four methods of *distribution*. In Figure 4, the average ratios of irrelevant passages are more than or close to 50% in documents with different relevance scores, which leads the *median* to a poor performance. These results state that people are more likely to make the document-level relevance judgment based on the overall relevance perception on passages within the document.

Compared with the minimum, maximum, median of passage-level relevance, other five weighted aggregation methods perform better on the nDCG and AUC metrics. The evidence of passage length on the importance of the passage is weaker than that of position and query similarity. It states that compared to the length of the passage, whether the passage content is similar to the query has greater influence on the importance of the passage. When considering the content of passages, the cosine similarity of passage content performs better than the exact match method. The Spearman’s correlation coefficient of *position decay* method is the greatest among all methods. We plot the distributions of the weighted aggregation of passage-level relevance within documents at different relevance levels according to *position decay* method as Figure 5 shows. We find that the aggregation results for the documents of different relevance levels differ in distribution. The aggregation results of more relevant documents tend to be higher.

### 4.3 Passage-level Relevance Sequence

Sequential passage-level relevance scores within a document may potentially affect the judgment of document-level relevance. The



**Figure 6: The relevance distributions of documents with and without the sub-sequences of passage-level relevance “000000” (left) and “33” (right).**

*distribution* and *aggregation* methods introduced above use independent passage-level relevance information. The relevance score sequence of sequential passages has not been taken into consideration. In this sub-section, we extract sub-sequences of passage-level relevance scores and analyse the relationship between sub-sequences and the document-level relevance.

We represent each document as a sequence composed of 0, 1, 2, and 3, which refer to the relevance scores of passages within the document. Then sub-sequences are extracted from the 1,050 relevance sequences by shifting a sliding window sized from 2 to 20 on the relevance sequences. We only reserve the sub-sequences that appear in at least 10% of documents because rare sub-sequences are not representative. Next, given the sub-sequences and document-level relevance, we analyse the most discriminating sub-sequences using the  $\chi^2$  test. A  $\chi^2$  score of 0 means the sub-sequence cannot discriminate among documents with different relevance judgments. 35 sub-sequences reject the null hypothesis with  $p\text{-value} < 0.001$ , such as the sub-sequences of “000000” and “33”. We plot the relevance distributions of documents with and without these two sub-sequences in Figure 6. We can see that in the documents with “000000”, 70.7% are irrelevant documents, only 9.3% are highly relevant documents. It shows that if a document contains “000000”, it is as highly possible that the document is irrelevant to the query. In all of the documents with “33”, the number of documents with “33” increases as the document-level relevance score increases, and 58.6% highly relevant documents contain the sub-sequence of “33”. It shows that the distributions of document-level relevance are different between the documents with and without specific sub-sequences, which may provide useful information for document-level relevance estimation.

To investigate how the sub-sequences perform in document-level relevance estimation, we frame a regression problem. Features are constructed based on the sub-sequences which reject the null hypothesis with  $p\text{-value} < 0.001$ . We use a binary label marking whether the sub-sequence appears in the relevance sequence of a document and use Gradient Boosting Decision Tree (GBDT) [7] with 5-fold cross-validation. Documents in one query is divided into one fold. The sub-sequences are extracted from the training set and applied on the test set. Based on the predicted document-level relevance, we also report the nDCG and the Spearman’s correlation coefficient between annotated and predicted document-level relevance in Table 4. We can see that the *sub-sequence* method can achieve better performance than all the four methods of *distribution*. It extra captures the signals of relevance score sequence of

**Table 5: The performances of different methods on estimating the document-level relevance. Queries are classified into *factual* and *intellectual* categories.**

| Category            | Method                                       | Spearman      |               | nDCG@5       |              | AUC          |              |
|---------------------|--|---------------|---------------|--------------|--------------|--------------|--------------|
|                     |  | Factual       | Intellectual  | Factual      | Intellectual | Factual      | Intellectual |
| <b>Distribution</b> | the fraction of irrelevant passages          | <b>-0.571</b> | <b>-0.563</b> | 0.779        | <b>0.765</b> | <b>0.841</b> | <b>0.777</b> |
|                     | the fraction of marginally relevant passages | 0.349         | 0.287         | 0.660        | 0.614        | 0.666        | 0.631        |
|                     | the fraction of fairly relevant passages     | 0.324         | 0.359         | 0.633        | 0.636        | 0.657        | 0.655        |
|                     | the fraction of highly relevant passages     | 0.519         | 0.509         | <b>0.785</b> | 0.750        | 0.781        | 0.720        |
| <b>Aggregation</b>  | minimum                                      | 0.041         | 0.042         | 0.511        | 0.557        | 0.531        | 0.504        |
|                     | maximum                                      | 0.581         | 0.593         | 0.783        | 0.736        | 0.841        | 0.776        |
|                     | median                                       | 0.414         | 0.435         | 0.676        | 0.679        | 0.724        | 0.678        |
|                     | mean   | 0.623         | <b>0.621</b>  | 0.843        | <b>0.814</b> | 0.865        | 0.798        |
|                     | position decay                               | <b>0.647</b>  | 0.586         | <b>0.853</b> | 0.805        | <b>0.872</b> | 0.792        |
|                     | passage length                               | 0.600         | 0.612         | 0.852        | 0.810        | 0.859        | <b>0.801</b> |
|                     | length with position decay                   | 0.623         | 0.599         | 0.851        | 0.786        | 0.866        | 0.798        |
|                     | exact match                                  | 0.590         | 0.526         | 0.840        | 0.762        | 0.845        | 0.751        |
| <b>Sequence</b>     | query similarity                             | 0.623         | 0.616         | 0.844        | 0.811        | 0.865        | 0.800        |
|                     | sub-sequence                                 | 0.550         | 0.514         | 0.806        | 0.735        | 0.868        | 0.745        |

continuous passages that single fraction of one type of passages can not capture. However, it performs worse than the *mean* method, probably because that the sub-sequences lose some information compared to the whole relevance sequence.

#### 4.4 Analysis on Query Types

Considering that the process of relevance judgment is affected by the search task types [20], we compare the performance of document-level relevance estimation methods introduced above between the documents in factual and intellectual queries in Table 5. We find that the method which has the best performance differs between these two query types. In factual queries, the *position decay* method performs better than the *mean* method. It states that passages within a document are not of the same importance in the document-level relevance judgment. Passages in top positions seem to be more important than those at the bottom. Taking the position information into consideration promotes the performance of document-level relevance estimation. However, in intellectual queries the *mean* method performs best among all the methods. The position decay assumption seems to be not suitable in this search scenario. It may be because that in factual queries, people are searching for objective facts such as the IELTS speaking test standard and ONE PIECE (see the examples in Table 1). They usually know the existence of the specific information that they are searching for and hope to find it more quickly. While in intellectual queries, the information people are looking for is usually subjective and narrative such as the reasons for the rise of oil price, they are uncertain about the existence of useful information and may be more patient in reading the documents from the top to the bottom. Therefore, the position of useful information has more influences on document-level relevance judgment in factual queries than intellectual ones.

#### 4.5 Summary

In this section, we use several methods based on the passage-level relevance to estimate the document-level relevance. Now we can

answer **RQ1** based on our findings in the experiment of document estimation methods:

- There is a strongly negative/positive correlation between the fraction of irrelevant/highly relevant passages and the document-level relevance.
- If there are more relevant passages in a document, the relevance score of the document is usually higher. People are more likely to make the document-level relevance judgment based on the overall relevance perception on passages within the document instead of the relevance of one certain passage.
- The position, length and query similarity of passages play different important roles in the document relevance judgment across different query types.
- Sequential irrelevant (or relevant) passages within a document potentially indicate an irrelevant (or relevant) judgment on the whole document.

### 5 DOCUMENT RANKING BASED ON PASSAGE-LEVEL RELEVANCE SIGNALS

In this section, we investigate whether the aggregation results of passage-level relevance can promote the performance of existing document retrieval models. We use BM25 to estimate the relevance of documents, which is a classic ranking algorithm with proven effectiveness in document ranking. It gives ranking scores for query-document pairs by counting the term frequency (TF) and inverse document frequency (IDF) of query terms appearing in documents. We calculate IDF on the whole THUCNews corpus. Based on it, we calculate the BM25 scores for query-document and query-passage pairs. Then the aggregated document-level relevance scores are obtained from the passage-level BM25 scores with the methods introduced in Section 4.2. We rerank the documents according to: 1) document-level BM25 scores; 2) aggregated relevance scores from the passage-level BM25 scores; 3) weighted summation of document-level and aggregated BM25 scores. We use  $\lambda$  to denote

**Table 6: Comparison of performances on document ranking between document-level relevance signals and aggregated passage-level relevance signals. Both document-level and passage-level relevance signals are calculated by BM25.**

| Method                   | $\lambda$ | nDCG@k       |              |              |
|--------------------------|-----------|--------------|--------------|--------------|
|                          |           | k=5          | k=10         | k=15         |
| BM25 <sub>document</sub> | 0         | 0.522        | 0.638        | 0.748        |
| minimum (CRD)            | 1         | 0.448        | 0.561        | 0.701        |
| maximum (DRD)            | 1         | <b>0.581</b> | <b>0.679</b> | <b>0.783</b> |
| median (AR)              | 1         | 0.481        | 0.589        | 0.712        |
| mean (AR)                | 1         | 0.491        | 0.620        | 0.725        |
| position decay           | 1         | 0.518        | 0.638        | 0.742        |
| passage length           | 1         | 0.497        | 0.616        | 0.729        |
| length with decay        | 1         | 0.517        | 0.637        | 0.743        |
| exact match              | 1         | 0.548        | 0.653        | 0.764        |
| query similarity         | 1         | 0.487        | 0.618        | 0.724        |
| minimum (CRD)            | 0.13      | 0.530        | 0.639        | 0.750        |
| maximum (DRD)            | 0.39      | <b>0.604</b> | <b>0.688</b> | <b>0.793</b> |
| median (AR)              | 0.29      | 0.558        | 0.650        | 0.763        |
| mean (AR)                | 0.37      | 0.530        | 0.646        | 0.754        |
| position decay           | 0.43      | 0.537        | 0.652        | 0.756        |
| passage length           | 0.37      | 0.555        | 0.655        | 0.759        |
| length with decay        | 0.42      | 0.567        | 0.664        | 0.770        |
| exact match              | 0.47      | 0.562        | 0.665        | 0.769        |
| query similarity         | 0.38      | 0.542        | 0.647        | 0.757        |

the weight of aggregated BM25 scores. The estimated document-level relevance is defined as follows:

$$\tilde{r}_d = \lambda f(s_{p_1}, \dots, s_{p_n}) + (1 - \lambda)s_d$$

where  $f$  denotes the aggregation methods, while  $s_d$  and  $s_{p_i}$  denote the BM25 scores of the document and the  $i$ -th passage. The first method is setting  $\lambda = 0$ , which indicates that the estimated document-level relevance is determined by the BM25 score of the document, and the second method is setting  $\lambda = 1$ , i.e., the estimated document-level relevance equals to the aggregated BM25 score of passages. In the third method, we use 5-fold cross-validation to split the training set and test set. Then we tune the  $\lambda$  from 0 to 1 with a step size of 0.01 and choose the best performed  $\lambda$  on the training set. We report the average  $\lambda$  learned on training sets and the average performance on test sets in Table 6.

We first look into the performance of aggregation methods when  $\lambda = 1$ . Compared with document-level BM25 scores, *maximum* and *exact match* achieve better ranking performances, which indicates that BM25 can be improved at finer-grained level by aggregating passage-level matching signals. We also observe that the aggregation methods with the learned  $\lambda$  outperform other methods with the fixed  $\lambda = 0$  or 1. Among the methods with the learned  $\lambda$ , *maximum* performs the best when  $\lambda = 0.39$ , followed by *length with decay* whose  $\lambda$  is 0.42. The values of  $\lambda$  in the two best methods indicate that the aggregated passage-level scores have rather large effect in determining the final document-level relevance score. Two-tailed t-test is performed to detect significant difference. However, our dataset has only 70 queries, which limits the statistical power of the evaluation experiments between different methods. The performance differences between the BM25 and other methods are

unstable among different queries, which do not lead to significant results in the t-test either.

Now we are at the position to answer **RQ2**. In summary, we find that the aggregation of fine-grained, passage-level BM25 scores can improve the performance of BM25 in the document ranking task. It indicates that it's feasible and beneficial to take advantage of the relationship between document-level relevance and passage-level relevance in document ranking.

## 6 CONCLUSION AND FUTURE WORK

In this paper, we investigate the role of perceived passage-level relevance in the document-level relevance judgment. Our dataset is constructed with the queries from query logs of a commercial search engine and the documents from a widely-used news corpus. We collect the relevance annotations for both query-document and query-passage pairs. We conduct a thorough analysis of how passage-level relevance signals determine or influence the document-level relevance. Our experiment results show that there is a strongly correlation between the distribution of irrelevant/highly relevant passages and the document-level relevance. The position, length and query similarity of passages play differently important roles in the document relevance judgment across factual and intellectual queries. Sequential irrelevant/relevant passages within a document potentially indicate an irrelevant/relevant judgment on the whole document. Finally, we find that it is effective to improve document ranking performance by incorporating passage-level relevance with document-level relevance.

*Implications and limitations.* We provide a detailed analysis of how people perceive the document-level relevance from passage-level relevance in this paper. As there are strong correlations between the document-level relevance and passage-level relevance, it is beneficial to consider passage-level relevance signals in document ranking. We would also like to highlight some limitations of this work. Our findings are based on the Chinese news data, where the content of documents is well-organized and of high quality. Whether these findings are applicable to other datasets that are not composed of news documents remains further investigation. We only use some basic features of passages (i.e., the position, length, and query similarity) in the weighting methods. There may be other useful features in the aggregation of passage-level relevance. Furthermore, the field based BM25F is worth being tried on the dataset with title and content information. Another limitation we need to highlight is the small size of query set. However, it is time-consuming and expensive to collect relevance labels for a large scale corpus and our dataset is the largest Chinese dataset with annotated relevance of query-document and query-passage pairs.

*Future work.* Interesting directions for future work include incorporating more factors in the analysis of passage-level relevance aggregation (e.g., the context information). Another feasible future work is to discover behavior patterns during the document-level relevance judgment such as eye movement.

## 7 ACKNOWLEDGEMENTS

This work is supported by Natural Science Foundation of China (Grant No. 61622208, 61732008, 61532011) and the National Key Research and Development Program of China (2018YFC0831700).

## REFERENCES

- [1] Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning* 11, 23–581 (2010), 81.
- [2] James P. Callan. 1994. Passage-level Evidence in Document Retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)*. Springer-Verlag New York, Inc., New York, NY, USA, 302–310. [https://dl.acm.org/citation.cfm?id=188490.188589](http://dl.acm.org/citation.cfm?id=188490.188589)
- [3] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to Rank: From Pairwise Approach to Listwise Approach. In *Proceedings of the 24th international conference on Machine learning*. ACM, 129–136.
- [4] Yixing Fan, Jiafeng Guo, Yanyan Lan, Jun Xu, Chengxiang Zhai, and Xueqi Cheng. 2018. Modeling Diverse Relevance Patterns in Ad-hoc Retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. ACM, New York, NY, USA, 375–384. <https://doi.org/10.1145/3209978.3209980>
- [5] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.
- [6] Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. 2003. An Efficient Boosting Algorithm for Combining Preferences. *J. Mach. Learn. Res.* 4 (Dec. 2003), 933–969. [https://dl.acm.org/citation.cfm?id=945365.964285](http://dl.acm.org/citation.cfm?id=945365.964285)
- [7] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.
- [8] Fredric C. Gey. 1994. Inferring Probability of Relevance Using the Method of Logistic Regression. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)*. Springer-Verlag New York, Inc., New York, NY, USA, 222–231. [https://dl.acm.org/citation.cfm?id=188490.188560](http://dl.acm.org/citation.cfm?id=188490.188560)
- [9] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A Deep Relevance Matching Model for Ad-hoc Retrieval. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM '16)*. ACM, New York, NY, USA, 55–64. <https://doi.org/10.1145/2983323.2983769>
- [10] Marti A. Hearst. 1993. *TextTiling: A Quantitative Approach to Discourse*. Technical Report. Berkeley, CA, USA.
- [11] Marti A. Hearst and Christian Plaut. 1993. Subtopic Structuring for Full-length Document Access. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '93)*. ACM, New York, NY, USA, 59–68. <https://doi.org/10.1145/160688.160695>
- [12] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional Neural Network Architectures for Matching Natural Language Sentences. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14)*. MIT Press, Cambridge, MA, USA, 2042–2050. [https://dl.acm.org/citation.cfm?id=2969033.2969055](http://dl.acm.org/citation.cfm?id=2969033.2969055)
- [13] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM '13)*. ACM, New York, NY, USA, 2333–2338. <https://doi.org/10.1145/2505515.2505665>
- [14] Oana Inel, Giannis Haralabopoulos, Dan Li, Christophe Van Gysel, Zoltán Szlávik, Elena Simperl, Evangelos Kanoulas, and Lora Aroyo. 2018. Studying Topical Relevance with Evidence-based Crowdsourcing. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. ACM, New York, NY, USA, 1253–1262. <https://doi.org/10.1145/3269206.3271779>
- [15] Thorsten Joachims. 2002. Optimizing Search Engines Using Clickthrough Data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02)*. ACM, New York, NY, USA, 133–142. <https://doi.org/10.1145/775047.775067>
- [16] Marcin Kaszkiel and Justin Zobel. 1997. Passage Retrieval Revisited. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '97)*. ACM, New York, NY, USA, 178–185. <https://doi.org/10.1145/258525.258561>
- [17] Marcin Kaszkiel and Justin Zobel. 2001. Effective ranking with arbitrary passages. *Journal of the American Society for Information Science and Technology* 52, 4 (2001), 344–364.
- [18] K Kong, R Luk, K Ho, and F Chung. 2004. Passage-based retrieval using parameterized fuzzy set operators. In *ACM SIGIR Workshop on Mathematical/Formal Methods for Information Retrieval*.
- [19] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. <http://www.jstor.org/stable/2529310>
- [20] Xiangsheng Li, Yiqun Liu, Jiaxin Mao, Zexue He, Min Zhang, and Shaoping Ma. 2018. Understanding Reading Attention Distribution During Relevance Judgement. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. ACM, New York, NY, USA, 733–742. <https://doi.org/10.1145/3269206.3271764>
- [21] Yuelin Li and Nicholas J. Belkin. 2008. A faceted approach to conceptualizing tasks in information seeking. *Information Processing and Management* 44, 6 (2008), 1822 – 1837. <https://doi.org/10.1016/j.ipm.2008.07.005> Adaptive Information Retrieval.
- [22] Xiaoyong Liu and W. Bruce Croft. 2002. Passage Retrieval Based on Language Models. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM '02)*. ACM, New York, NY, USA, 375–382. <https://doi.org/10.1145/584792.584854>
- [23] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. In *HCOMP*.
- [24] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 3111–3119.
- [25] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Jingfang Xu, and Xueqi Cheng. 2017. DeepRank: A New Deep Architecture for Relevance Ranking in Information Retrieval. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. ACM, New York, NY, USA, 257–266. <https://doi.org/10.1145/3132847.3132914>
- [26] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (April 2009), 333–389. <https://doi.org/10.1561/1500000019>
- [27] Stephen Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple BM25 Extension to Multiple Weighted Fields. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management (CIKM '04)*. ACM, New York, NY, USA, 42–49. <https://doi.org/10.1145/1031171.1031181>
- [28] S. E. Robertson and S. Walker. 1994. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)*. Springer-Verlag New York, Inc., New York, NY, USA, 232–241. [https://dl.acm.org/citation.cfm?id=188490.188561](http://dl.acm.org/citation.cfm?id=188490.188561)
- [29] Gerard Salton, J. Allan, and Chris Buckley. 1993. 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 (SIGIR '93)*. ACM, New York, NY, USA, 49–58. <https://doi.org/10.1145/160688.160693>
- [30] Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. Learning Semantic Representations Using Convolutional Neural Networks for Web Search. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)*. ACM, New York, NY, USA, 373–374. <https://doi.org/10.1145/2567948.2577348>
- [31] Andrew Trotman, Nils Pharo, and Dylan Jenkinson. 2007. Can we at least agree on something. In *Proceedings of the SIGIR 2007 Workshop on Focused Retrieval*. 49–56.
- [32] Mengqiu Wang and Luo Si. 2008. Discriminative probabilistic models for passage based retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 419–426.
- [33] Ryen W. White, Ian Ruthven, and Joemon M. Jose. 2002. Finding Relevant Documents Using Top Ranking Sentences: An Evaluation of Two Alternative Schemes. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '02)*. ACM, New York, NY, USA, 57–64. <https://doi.org/10.1145/564376.564389>
- [34] Ross Wilkinson. 1994. Effective Retrieval of Structured Documents. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)*. Springer-Verlag New York, Inc., New York, NY, USA, 311–317. [https://dl.acm.org/citation.cfm?id=188490.188591](http://dl.acm.org/citation.cfm?id=188490.188591)
- [35] H. C. Wu, Robert W. P. Luk, K. F. Wong, and K. L. Kwok. 2007. A Retrospective Study of a Hybrid Document-context Based Retrieval Model. *Inf. Process. Manage.* 43, 5 (Sept. 2007), 1308–1331. <https://doi.org/10.1016/j.ipm.2006.10.009>
- [36] Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. 2008. Interpreting TF-IDF Term Weights As Making Relevance Decisions. *ACM Trans. Inf. Syst.* 26, 3, Article 13 (June 2008), 37 pages. <https://doi.org/10.1145/1361684.1361686>
- [37] Wensi Xi, Richard Xu-Rong, Christopher SG Khoo, and Ee-Peng Lim. 2001. Incorporating window-based passage-level evidence in document retrieval. *Journal of information science* 27, 2 (2001), 73–80.
- [38] Jun Xu and Hang Li. 2007. AdaRank: A Boosting Algorithm for Information Retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*. ACM, New York, NY, USA, 391–398. <https://doi.org/10.1145/1277741.1277809>
- [39] Nikita Zhiltsov, Alexander Kotov, and Fedor Nikolaev. 2015. Fielded Sequential Dependence Model for Ad-Hoc Entity Retrieval in the Web of Data. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. ACM, New York, NY, USA, 253–262. <https://doi.org/10.1145/2766462.2767756>