Investigating Reading Behavior in Fine-grained Relevance Judgment

Zhijing Wu, Jiaxin Mao, Yiqun Liu*, Min Zhang, and 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
wuzhijing.joyce@gmail.com,yiqunliu@tsinghua.edu.cn

ABSTRACT

A better understanding of users' reading behavior helps improve many information retrieval (IR) tasks, such as relevance estimation and document ranking. Existing research has already leveraged eye movement information to investigate user's reading process during document-level relevance judgments and the findings were adopted to build more effective ranking models. Recently, fine-grained (e.g., passage or sentence level) relevance judgments have been paid much attention to with the requirements in conversational search and QA systems. However, there is still a lack of thorough investigation on user's reading behavior during these kinds of interaction processes. To shed light on this research question, we investigate how users allocate their attention to passages of a document during the relevance judgment process. With the eye-tracking data collected in a laboratory study, we show that users pay more attention to the "key" passages which contain key useful information. Users tend to revisit these key passages several times to accumulate and verify the gathered information. With both content and user behavior features, we find that key passages can be predicted with supervised learning. We believe that this work contributes to better understanding users' reading behavior and may provide more explainability for relevance estimation.

KEYWORDS

eye-tracking, passage-level cumulative gain, relevance judgment

ACM Reference Format:
Zhijing Wu, Jiaxin Mao, Yiqun Liu*, Min Zhang, and Shaoping Ma. 2020. Investigating Reading Behavior in Fine-grained Relevance Judgment. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401305

1 INTRODUCTION

Users' reading behavior is one of the most important signals for understanding the information seeking process. A number of studies [1, 3, 5, 12] have investigated the reading behavior patterns based on users' eye movements (i.e., sequences of fixations and saccades). It has been demonstrated that the eye movements are useful in understanding users' search process [1], generating implicit feedback for relevance estimation [3] and improving the performance of machine reading comprehension models [12]. Relevance judgment is essential for the evaluation of search systems. Most of existing works focused on the document-level relevance judgment. For example, Li et al. [5] investigated users' reading behavior patterns on documents during the relevance judgment process. Based on the heuristics summarized from these reading behavior patterns, they further proposed a Reading Inspired Model (RIM), which improved the performance and explainability in the document ranking task [6]. Nowadays, conversational search and question answering systems become popular and such users' need usually can be satisfied by a single passage or even a sentence. To improve performance in these tasks, we need to investigate fine-grained relevance judgment. However, little is known about the relationship between the reading behavior and the fine-grained relevance judgment (e.g., the passage-level or sentence-level relevance).

As relevant content could be located at any position of a Web document, we consider that a finer-grained relevance judgment helps better capture these local relevance signals. Previous work has attempted to improve ranking performance by introducing fine-grained relevance signals [2, 8]. Recently, Wu et al. [10] proposed Passage-level Cumulative Gain (PCG), which represents how useful information accumulates passage by passage when a user sequentially reads a document. The context-aware PCG avoids the need to formally split a document into independent passages and successfully improves the performance of document ranking task. However, there is a lack of analysis between users' reading behavior and passage-level cumulative gain.

In this paper, we investigate users' reading behavior during the relevance judgment process at the passage level. When a user reads a document with an information need, some useful passages play a key role in the relevance perception process. Therefore, we analyze users' reading behavior patterns on these key passages. Furthermore, we try to utilize the patterns we found to predict which passages are key ones. Our research questions as follows:

- **RQ1:** What is the relationship between users' reading behavior and perceived fine-grained relevance?
- **RQ2:** Can we predict key passages with user behavior features?

To shed light on these research questions, we use the eye-tracking dataset [1] collected by Li et al. [5] in a laboratory user study. Then we

[1] http://www.thuir.cn/group/~YQLiu/

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401305

Table 1: Statistics of our dataset.

#Tasks	#Documents	#Passages	#PCG annotations
14	56	513	1,539

collect annotations of passage-level cumulative gain for documents in this dataset according to Wu et al. [10]. To answer RQ1, we analyze the distribution and transition of eye movements. Furthermore, we define key passage prediction as a binary classification task and demonstrate the effectiveness of features extracted from implicit and explicit feedback of users.

The remainder of this paper is organized as follows. We review related work in Section 2. Then in Section 3, we introduce the dataset we used. Section 4 describes the reading behavior patterns and the prediction model to address our research questions. In Section 5, we conclude this paper and suggest directions for future research of this work.

2 RELATED WORK

There are many existing studies focusing on users' reading behavior. Buscher et al. [1] examined the relationship between eye movement measures and user-perceived relevance. They further showed that gaze-based feedback is very useful in improving the quality of Web search. Gwizdka [3] utilized eye movement patterns to better understand the cognitive processing of text documents at different degrees of relevance. Zheng et al. [12] investigated how human reads and allocates their attention during reading comprehension processes. Based on features extracted from user behavior, they significantly improved the performance of MRC models. Recently, Li et al. [6] improved the performance and explainability in the document ranking task based on users' reading behavior patterns summarized by Li et al. [5]. These studies reveal the potential and wide application of reading behavior.

Fine-grained relevance judgments have also drawn much attention in recent studies. [11] provided a detailed analysis of how passage-level relevance signals determine the relevance judgment of the whole document. Wu et al. [10] proposed context-aware passage-level cumulative gain and demonstrated its effectiveness in improving the performance of document ranking task. Compared to these studies, our work focuses on the relationship between users' reading behavior and fine-grained relevance judgments.

3 DATASET

We use the eye-tracking dataset collected by Li et al. [5] in a laboratory user study, which includes search tasks, documents, eye-tracking data, and highlighted text. There are 15 search tasks and 60 documents in this dataset. Participants were required to highlight the relevant texts and make relevance judgment for documents with respect to the corresponding search intent. Their eye movements during reading the documents were recorded by an eye-tracker. Each document was annotated by 7 or 8 participants.

We further collect annotations of the passage-level cumulative gain (PCG) for each document according to Wu et al. [10]. We use a four-grade PCG judgment scale to annotate the PCG labels (0: no gain, 1: low gain, 2: moderate gain, 3: high gain). PCG labels of a document d can be described as a sequence $G_d = \{g_1, g_2, ..., g_n\}$,

Table 2: Distributions of document-level cumulative gain (DLCG) and passage-level cumulative gain (PCG). The Avg. #P and Avg. #W mean the average number of passages and words within documents.

Document-level				Passage-level	
DLCG	**Proportion**	**Avg. #P**	**Avg. #W**	**PCG**	**Proportion**
0	0.214	9.6	466	0	0.402
1	0.196	7.9	532	1	0.193
2	0.089	9.0	508	2	0.140
3	0.500	9.5	464	3	0.265

where n is the number of passages in d and g_i denotes the degree of gain that the user obtains from the first i passages of d. Therefore, g_n denotes the whole-document-level cumulative gain (DLCG) of d. We regard the i-th passage as a "key" passage if it satisfies $g_i > g_{i-1}$, which indicates that this passage contains key useful information and users' perceived gain increases after they read this passage. Noted that we set g_0 to 0. The PCG annotations avoid the problem of how to aggregate relevance scores of independent passages to get document-level relevance. In this work, one search task (i.e., please find the story introduction of the film "Flirting Scholar") is excluded because we find that it's difficult to annotate the PCG in this search task. A paragraph is taken as a passage. For each document, we obtain three PCG sequences from three different annotators. We use the majority vote of the three annotators as the final PCG label. Krippendorff's α [4] for ordinal data is used to measure the inter-person agreement of PCG annotations, which is 0.844, indicating an almost perfect agreement level.

Statistics of the dataset[2] are shown in Table 1. Table 2 shows the distributions of DLCG and PCG annotations. 50% of documents in the dataset fully satisfy the information needs (i.e., DLCG = 3). 40.2% of passages contain no useful information, while 26.5% of passages fully satisfy the information needs of users. In the dataset, 17% of passages are key passages, meaning that users' perceived information gain gets increased after reading them.

4 RESULTS

4.1 Analysis on Reading Behavior

To answer **RQ1**, we analyze the relationship between users' reading behavior and the passage-level cumulative gain by examining the fixation distribution and examination sequence. We also analyze users' explicit feedback (i.e., highlighting behavior) to better understand the relationship.

4.1.1 Fixation Distribution. Based on the eye-tracking data, we analyze users' eye fixations, during which users' eyes land on an object and remain relatively stationary for a brief period of time. The number and duration of fixations have been regarded as effective implicit feedback for improving document ranking. We calculate the average number and duration of fixations on each word within the non-key and key passages, respectively. Their distributions are shown in Figure 1. In non-key passages, there is 0.18 fixation on each word on average. The total fixation duration per word is 55.4 milliseconds. In key passages, the average number of fixations is

[2]The PCG labels for this dataset is now available at http://www.thuir.cn/group/~YQLiu/

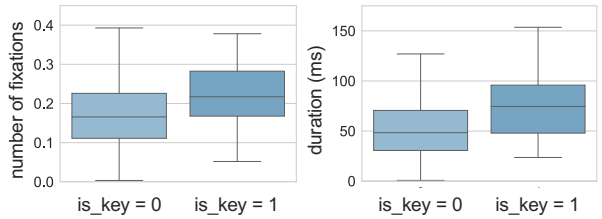

Figure 1: The average number of fixations and fixation duration on each word within non-key (is_key=0) and key (is_key=1) passages. The differences between the non-key and key passages are statistically significant at $p < 0.01$, using two-tailed t-test.

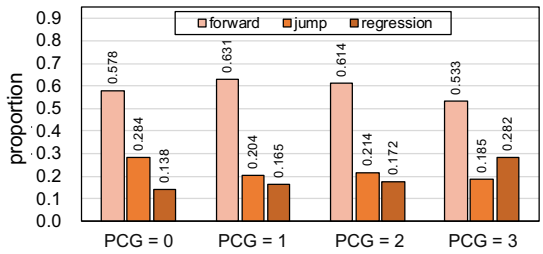

Figure 2: Distributions of reading transitions among passages with different PCG labels.

0.23 and the duration is 77.5 milliseconds. An independent two-tailed t-test is performed to detect the significance of differences between non-key and key passages and shows that their differences are both statistically significant. The above observations illustrate that key passages attract more attention from users during the relevance judgment process. Users obtain useful information by reading these passages and their perceived gain gets increased.

4.1.2 Examination Sequence. To further understand the temporal sequence of users' reading behavior, We analyze the fixation transitions among passages. We divide the transition behavior into three categories according to McDonald and Shillcock [7]: 1) Forward: users' fixations go to the next passage; 2) Regression: users' fixations go to the previous passages 3) Skip: users skip some passages and read posterior passages. Figure 2 shows the distributions of reading transitions. For example, "0.578" in the figure means that after users read the no-gain passages (i.e., PCG=0), 57.8% of the transitions are forward transition. We observe that when the information need is not fully satisfied (i.e., PCG<3), users tend to continue reading new passages to accumulate more useful information. As the cumulative gain increases, users may want to verify the gathered information, so they revisit previous passages more frequently. When they get enough useful information (i.e., PCG=3), 28.2% of the fixation transitions are regression transition.

We further analyze the transition behavior with key passages. We calculate the proportion of key passages in passages which are skipped by users, and in passages which are examined by users through forward, jump, and regression reading. The results are shown in Figure 4. For example, "0.844" in the figure means that during the reading process, 84.4% of the passages skipped by users are non-key passages. We find that the proportion of key passages

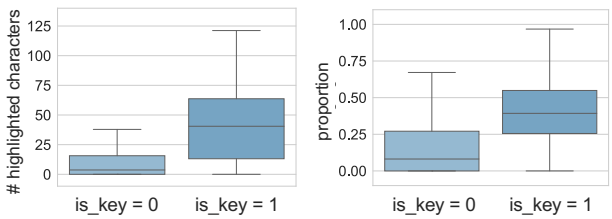

Figure 3: The average number and proportion of highlighted Chinese characters in non-key (is_key=0) and key (is_key=1) passages. The differences between the non-key and key passages are statistically significant at p < 0.01 (two-tailed t-test)

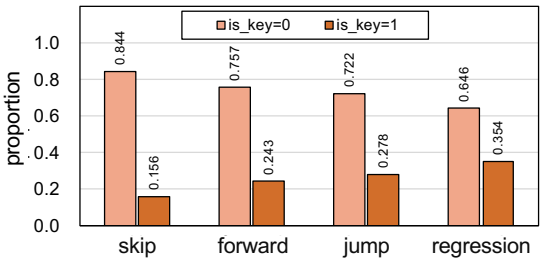

Figure 4: Distributions of key passages in passages skipped by users, and in passages examined by users through forward, jump, and regression reading.

(i.e., 15.6%) in skipped passages is slightly lower than that in the document (i.e., 17%). However, the proportions of key passages in examined passages are all higher than 17%. During the reading process, key passages are revisited more times than skipped. We explain that why users revisit key passages more frequently than non-key passage is to verify the useful information.

4.1.3 Highlighting Behavior. In our dataset, after participants made the relevance judgment in the lab study of, they were asked to highlight the relevant parts of text that were helpful for the search task. To understand the relationship between users' explicit feedback of relevant parts and key passages, we compare the average number and proportion of highlighted Chinese characters in non-key and key passages, as shown in Figure 3. In the non-key passages, users highlighted 13.8 characters as relevant text on average, which are 16.8% of all characters. In the key passages, the average number of highlighted characters is 45.9 and the proportion is 39.8%, which is higher than that in non-key passages. The differences between non-key and key passages are both statistically significant using an independent two-tailed t-test. It shows that users find more useful information in key passages.

4.1.4 Summary. In this section, we investigate the relationship between users' reading behavior and fine-grained relevance judgment. To answer **RQ1**, our findings are as follows: 1) Users pay more attention to and highlight more relevant text in the key passages, which contain key useful information; 2) Users tend to continue reading new passages to accumulate more useful information when they are not fully satisfied; 3) Users choose to revisit key passages frequently to accumulate and verify the gathered information.

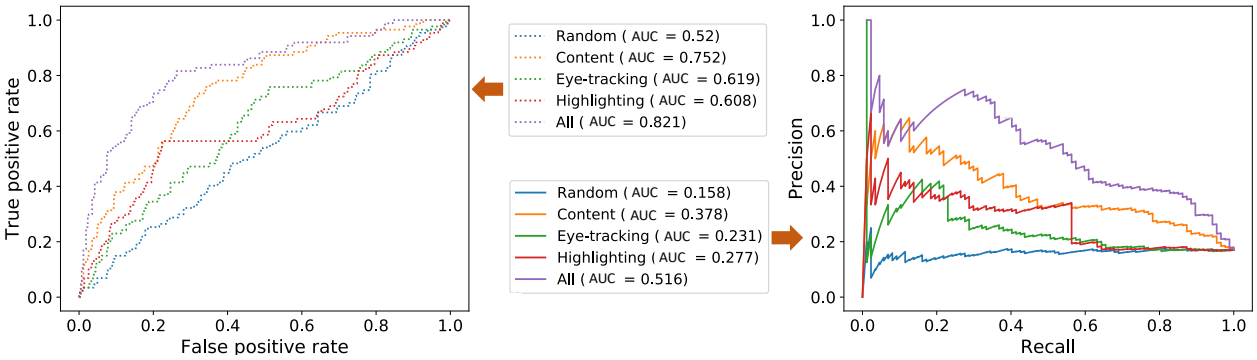

Figure 5: The ROC curve (left) and precision-recall curve (right) of key passage prediction results.

4.2 Key Passage Prediction

Since key passages play a key role during the relevance judgment process, we further try to predict key passages with both content and user behavior. We define the prediction task as a binary classification task and use the AUC of the ROC curve and precision-recall curve to evaluate the performance.

Three groups of features are used for the prediction: content, eye-tracking and highlighting. For content features, we extract eight features for each passage according to Qin and Liu [9], including the number of words in the passage, the average TF, IDF, and TF*IDF values of query terms in the passage, scores of BM25 and three language models. We further use the eight-dimensional vector to represent each passage using the bag-of-words model and calculate the cosine similarity between p_i and p_j $(1 \leq j < i)$. For the prediction of i-th passage p_i, the input content features include the eight features extracted from P_i, as well as the maximum, minimum and mean values of the eight features extracted from the similarity between the i-th passage and the first $i-1$ passages. The eye-tracking features include the number of fixations, fixation duration, and the proportions of the three types of transitions in the reading process. The highlighting features include the number and proportion of highlighted Chinese characters within the passage.

We use Gradient Boosting Classifier and perform 5-fold cross-validation for the prediction task. Results are shown in Figure 5. We observe that all of the three groups of features are useful for the key passage prediction. Combining all the features further improves the performance. With both content and user behavior features, the AUC scores of the ROC and precision-recall curve reach 0.821 and 0.516, respectively. Even only using the content features, the prediction model achieves a ROC-AUC score of 0.752. It shows that key passages can be predicted with supervised learning.

5 CONCLUSION

In this paper, we mainly investigate users' reading behavior during the relevance judgment process and link the behavior to fine-grained relevance. We find that key passages that contain key useful information attract more user attention. When the information need is not fully satisfied, users tend to continue reading new passages. During the whole reading process, users revisit key passages more frequently than skip to accumulate and verify the gathered information. We further show that key passages can be predicted with the content and user behavior features. Our work is the first research to analyze the relationship between reading behavior and context-aware fine-grained relevance judgments. As for future work, we would like to investigate users' reading behavior when they are seeking useful information for a search task among multiple retrieved documents. We also plan to study the relationship between the passage-level feedback and some implicit feedbacks such as mouse scrolling, mouse movement, or viewport features. We believe that a deeper understanding of the passage-level feedback can further help improve the relevance estimation.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (2018YFC0831700), Natural Science Foundation of China (Grant No. 61732008, 61532011, 61902209) and Beijing Academy of Artificial Intelligence (BAAI).

REFERENCES

[1] Georg Buscher, Andreas Dengel, Ralf Biedert, and Ludger van Elst. 2012. Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond. *ACM Transactions on Interactive Intelligent Systems* 1 (01 2012), 9.
[2] Yixing Fan, Jiafeng Guo, Yanyan Lan, Jun Xu, Chengxiang Zhai, and Xueqi Cheng. 2018. Modeling Diverse Relevance Patterns in Ad-hoc Retrieval. In *SIGIR '18* (Ann Arbor, MI, USA). ACM, New York, NY, USA, 375–384.
[3] Jacek Gwizdka. 2014. Characterizing Relevance with Eye-Tracking Measures. In *Proceedings of the 5th Information Interaction in Context Symposium* (Regensburg, Germany) *(IIiX'14)*. Association for Computing Machinery, NY, USA, 58–67.
[4] Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures* 1, 1 (2007), 77–89.
[5] Xiangsheng Li, Yiqun Liu, Jiaxin Mao, Zexue He, Min Zhang, and Shaoping Ma. 2018. Understanding Reading Attention Distribution during Relevance Judgement. In *CIKM'18*. 733–742.
[6] Xiangsheng Li, Jiaxin Mao, Chao Wang, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Teach Machine How to Read: Reading Behavior Inspired Relevance Estimation. In *SIGIR'19*. 795–804.
[7] Scott A. McDonald and Richard C. Shillcock. 2003. Low-level predictive inference in reading: the influence of transitional probabilities on eye movements. *Vision Research* 43, 16 (2003), 1735 – 1751.
[8] 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 *CIKM '17* (Singapore). ACM, New York, NY, USA, 257–266.
[9] Tao Qin and Tie-Yan Liu. 2013. Introducing LETOR 4.0 Datasets. *CoRR* abs/1306.2597 (2013). arXiv:1306.2597 http://arxiv.org/abs/1306.2597
[10] Zhijing Wu, Jiaxin Mao, Yiqun Liu, Jingtao Zhan, Yukun Zheng, Min Zhang, and Shaoping Ma. 2020. Leveraging Passage-level Cumulative Gain for Document Ranking. In *The World Wide Web Conference*.
[11] Zhijing Wu, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Investigating Passage-Level Relevance and Its Role in Document-Level Relevance Judgment. In *SIGIR'19* (Paris, France). 605–614.
[12] Yukun Zheng, Jiaxin Mao, Yiqun Liu, Zixin Ye, Min Zhang, and Shaoping Ma. 2019. Human Behavior Inspired Machine Reading Comprehension. In *SIGIR'19*. New York, NY, USA, 425–434.