Understanding Reading Attention Distribution during Relevance Judgement

Xiangsheng Li†, Yiqun Liu†*, Jiaxin Mao†, Zexue He‡, 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
‡Beijing Normal University, Beijing, China
yiqunliu@tsinghua.edu.cn

ABSTRACT

Reading is a complex cognitive activity in many information retrieval related scenarios, such as relevance judgement and question answering. There exists plenty of works which model these processes as a matching problem, which focuses on how to estimate the relevance score between a document and a query. However, little is known about what happened during the reading process, i.e., how users allocate their attention while reading a document during a specific information retrieval task. We believe that a better understanding of this process can help us design better weighting functions inside the document and contributes to the improvement of ranking performance. In this paper, we focus on the reading process during relevance judgement task. We designed a lab-based user study to investigate human reading patterns in assessing a document, where users' eye movements and their labeled relevant text were collected, respectively. Through a systematic analysis into the collected data, we propose a two-stage reading model which consists of a preliminary relevance judgement stage (Stage 1) and a reading with preliminary relevance stage (Stage 2). In addition, we investigate how different behavior biases affect users' reading behaviors in these two stages. Taking these biases into consideration, we further build prediction models for user's reading attention. Experiment results show that query independent features outperform query dependent features, which indicates that users allocate attentions based on many signals other than query terms in this process. Our study sheds light on the understanding of users' attention allocation during relevance judgement and provides implications for improving the design of existing ranking models.

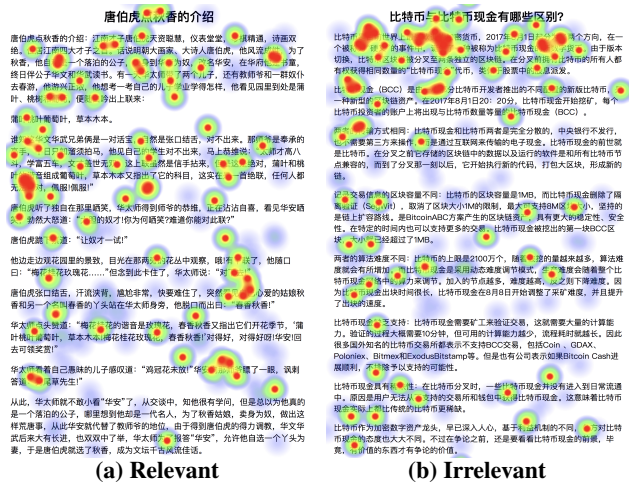
KEYWORDS

Attention; Relevance Judgement; User Behavior Analysis

ACM Reference format:
Xiangsheng Li†, Yiqun Liu†, Jiaxin Mao†, Zexue He‡, Min Zhang†, Shaoping Ma†. 2018. Understanding Reading Attention Distribution during Relevance Judgement. In *Proceedings of The 27th ACM International Conference on Information and Knowledge Management, Torino, Italy, October 22–26, 2018 (CIKM '18),* 10 pages.
DOI: 10.1145/3269206.3271764

CIKM '18, Torino, Italy

DOI: 10.1145/3269206.3271764



(a) Relevant **(b) Irrelevant**

Figure 1: Users' fixation distributions while reading a relevant document (a) and an irrelevant document (b). For both documents, users pay much more attention to the top part of document than the bottom part. This position bias is more apparent in the irrelevant document.

1 INTRODUCTION

Reading is a complex cognitive activity that involves the orchestration of many different stages of information processing [36]. Eye movement, which is composed of a sequence of fixations and saccades, are found to be useful in analyzing human's reading behavior. During a fixation, the eyes land on an object and remain relatively stationary for a brief period of time (typically 200-250 msec). A saccade is the rapid eye movement between fixations to move the eye-gaze from one point to another, typically lasting 20-50 msec [36]. In cognitive psychology, a number of computational models, such as EZ-Reader [35, 36], SWIFT [15, 16], and the Bayesian reading model [5], have been proposed to account for the reading behavior. These models provide insights into the understanding of individual's general reading behavior.

Recently, understanding users' reading behavior during their information seeking process has drawn much attention in IR-related studies. Eye tracking, as an unobtrusive and precise measure of user's visual attention, is utilized to investigate user's cognitive process during the search process [8], as well as to generate implicit feedback for relevance [18] and text quality [30]. Because of the

*Corresponding author

existence of information needs, the reading behavior in information retrieval is often inconsistent with general reading behaviors. Two examples of users' reading process while performing relevance judgment process are shown in Figure 1. We can observe that users' reading attention has strong position bias in vertical positions and the relevance of the document has an impact on users' reading behavior. Besides, users' reading attention during the search is also influenced by other factors, such as search task types [12] and query terms [41]. Therefore, general reading models established in cognitive psychology are not necessarily effective in explaining users' behavior in IR tasks.

In this paper, we focus on investigating the reading behavior in relevance judgment task, where the assessor will read a document and judge whether the document is relevant or not according to the current query or search task. As relevance judgment is essential for the evaluation of search systems, it is important to understand how the assessors read the document and make judgement. Previous studies have shown that the cognitive process during relevance judgment is highly complex [4] and proposed a series of assumptions to model the corresponding reading behavior. For example, Wu et al. [41] proposed a strong *query-centric assumption*, which assumes that the relevant information only locates in the contexts around the query terms. Recently, researchers also adopt attention mechanism while performing machine reading with deep neural networks [3, 43]. It simulates human's reading behavior with flexible attention on the important part of the content. However, these approaches are not based on observations of users' reading behavior. To construct better computational models for relevance judgment (i.e., retrieval models), we need to better understand human's reading patterns and attention distribution in real relevance judgment scenarios.

Since eye movement is tightly coupled with cognitive attention during reading in our brains [26] and may serve as a measurable indicator of the reading process, we design a dedicated user study to simulate the relevance judgment scenario and use an eye-tracker to collect participants' eye-movements during the completion of relevance judgment tasks. For each relevance judgment task, the participants are required to read a document and label its relevance according to a specific information need. Then, the same document will be presented again and the participants need to highlight the relevant text that is helpful for the task. Based on the collected data, our study aims to address the following research questions:

- **RQ 1**: How does user make the relevance judgment while reading a document?
- **RQ 2**: How does user allocate his/her attention during the relevance judgment process?
- **RQ 3**: What are the factors that affect user's attention allocation in relevance judgment?

Through analyzing users' eye movements and their labeled relevant text, we found that users exhibit similar reading patterns at the beginning part of the document (about top 20% to 40%). However, their behaviors become significantly different on the later parts of relevant and irrelevant documents. This is probably because that users perceive a preliminary relevance and this perception influences their follow-up reading behavior. Thus, we propose a two-stage reading model which consists of: 1) a preliminary relevance judgement stage, and 2) a reading with preliminary relevance stage. In the first stage, users tend to read the text carefully until they form a preliminary relevance judgement in their mind. In the second stage, users will have different reading behaviors with different preliminary judgements in the first stage. They will gather evidence either to acquire knowledge or to validate the judgement. During this two-stage relevance judgment process, the reading behavior is influenced by a number of factors such as position, linguistic features, search intents, and query terms. These factors have different impacts on users' reading behavior. It suggests that reading is a complex cognitive process and users rely on different signals to allocate their attention in different stages. As attention allocation is one of the most fundamental cognitive mechanisms of human being [25], our findings can be regarded as an attempt to explain how this mechanism works in relevance judgement task. Besides, understanding the influencing factors on attention can also benefit the design of retrieval models.

The remainder of this paper is organized as follows. In Section 2, we review some related studies to our work. Section 3 describes the design of our study and the data collected in our study. In Section 4, we analyze user reading process and propose the two-stage reading model, which addresses **RQ1** and **RQ2**. To investigate **RQ3**, we analyze reading behavior biases in Section 5 and build prediction models for attention in Section 6. Finally, we conclude our work and propose future research directions in Section 7.

2 RELATED WORK

2.1 Reading Model

Reading is an essential process for acquiring information during primarily textual information search. Based on users' eye movements, the reading patterns and further how language is processed can be understood [36]. There exists a number of reading models elaborating the information acquisition during the reading process. The EZ Reader model is a cognitively-controlled, serial-attention model of reading eye movements [35, 36]. It proposed different cognitive stages that consider word identification, visual processing, attention, and control of the oculomotor system as joint determinants of eye movement in the reading process. Specifically, it discovered that fixation and saccade alternate because saccade programming has a labile stage. If the next word is recognized during this labile stage, users will progressively move to the next word instead of programming currently attended word. Rayner et al [34] further showed that users are able to identify a skimmed word by using *parafoveal preview*. Thus, the word of which words to skip not only relies on the prior context but also a preview of the word itself [19]. There are other factors that are shown to influence word recognition such as word length [22], orthographic features, the predictability of the word in the context [13] and morphological features [14].

These findings illustrate the information acquisition during the reading process. However, the reading patterns may be different in information retrieval tasks since there are other factors such as query terms [41] and search task types [10]. The cognitive process in information search remains to be further investigated.

2.2 Machine Reading

Human reading inspires machine reading developed in many information retrieval tasks such as relevance judgement [33] and question answering [38]. Systems are taught to simulate human reading for a specific search task. A commonly used strategy is to compute a matching score (often called relevance) between the query and the document. To better build systems that are able to perform machine

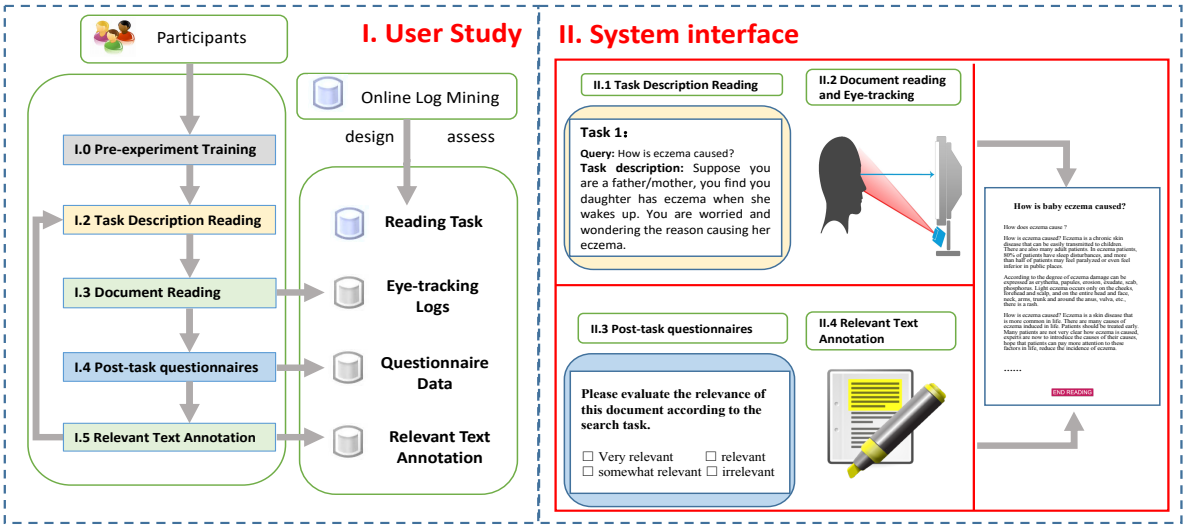


Figure 2: User study procedure. The system interface is translated from Chinese.

reading in information retrieval tasks, researchers proposed a number of approaches based on human reading behavior. Wu et al [41] proposed the *query-centric assumption*, which strongly assumes that the relevant information for a query only locates in the contexts around query terms. Wu et al [42] applied a visual semantic attention to extract keywords for document classification. Bidirectional Encoding [38] is utilized to process document in two directions in terms of *parafoveal preview* [34]. Another popular method applied in deep learning is attention mechanism [3]. It simulates human reading behavior with flexible attention on the important part of context information. However, these methods have not been validated whether they match with human reading patterns. Human reading attention in information retrieval is thus important and necessary to be investigated.

2.3 Relevance Judgement

Relevance is a central notion in information retrieval. Users are involved to judge the relevance degree of documents with respect to a given query. It has been contended that the relevance of an information object to the information need of a specific user is a subjective and multi-dimensional concept, which encompasses various properties and characteristics of the sought information objects [6]. Zhang et al [44] showed some evidence of quantum-like cognitive interference in human decision making and proposed a dynamic model for cognitive interference, which leads to a better modeling and explanation of user behaviors in relevance judgement process. In addition, users express their information needs by issuing a sequence of query terms, which are likely to be ambiguous and noisy [21]. This makes us difficult to predict the relevance between the document and the issued queries. To better mode the process of relevance judgement, a number of relevance feedback techniques have been introduced in the literature, which vary from explicit to implicit.

Explicit relevance feedback is collected by users annotation, but it usually contains more noise compared to implicit feedback [1]. Though it is noisy, it allows us to collect several distinct signals of the users interests and satisfactions. Loboda et al [27] recruited participants to mark relevant text to model word-level relevance, which explicitly represents the relevant content in a document. As for implicit feedback, it is more close to users' real perception compared to explicit feedback. Over the last few years, affective and physiological features have been considered as implicit feedback techniques [2, 32], such as mouse movement [9] and eye movement [7, 8]. In this paper, we construct both explicit and implicit feedback to model reading attention by collecting labeled relevant text and eye movement, respectively. The collected reading attention is helpful to understand the cognitive process in relevance judgement.

Table 1: Examples of user study tasks.

Query	Domain	Search task type	Background
What is TPP?	Politics	Intellectual	On October 5, 2015, the basic agreement was reached through negotiations on the Trans-Pacific Partnership (TPP). You want to understand what TPP is?
How is eczema cause?	Health	Factual	Assuming you are the parent of a child, you find that your baby has eczema after waking up. You want to find out what caused the baby's eczema.
Immigrate abroad	Culture	Intellectual	You are confused about whether or not to immigrate abroad and wondering what benefits and disadvantages immigrants may bring you, so you want to listen to more people's opinions on the Internet.

3 DATA COLLECTION

In this section, we describe the settings of our user study and the dataset we collected.

3.1 Tasks

We designed a laboratory user study to collect participants' eye movement during relevance judgment. Participants were required to make relevance judgment for a series of documents with respect to the corresponding search tasks. We chose 15 queries from the NTCIR-13 We Want Web (WWW) task [29]. For each query, we created a background story to describe the corresponding search task. The selected search tasks came from three domains: Politics, Health and Culture, and cover both *factual* and *intellectual* product aspects defined by Li and Belkin [24] (7 tasks were factual and 8 were intellectual). We sampled 4 documents for each search task and had 60 documents in total. In the user study, our laboratory system would randomly present one document for each search task to a participant. Therefore, each participant was required to read 15 documents and each document was read by 7-8 users. Note that the language used in our study is Chinese, i.e., all the task descriptions, search systems, and instructions are in Chinese. Examples of the translated queries and task descriptions are shown in Table 1.

3.2 Participants

We recruited 29 university students via email and online social networks to take part in our user study. The ages of participants ranged from 17 to 28 and their majors included humanities, social science, arts, and engineering. All the participants are native Chinese speakers and have college-level Chinese reading and writing skills. To assure the validity of collected eye movements, we required all participants to possess normal corrected eyesight with correction (including astigmatism and strabismus). It took about one hour to complete the user study and we paid each participant about US$15 after they completed all the tasks.

3.3 Procedure

The procedure of the user study is shown in Figure 2. To make sure that each participant was familiar with the experimental procedure, an example task was used as a tutorial in the pre-experiment training stage. After the pre-experiment training stage, they were asked to complete all 15 search tasks. The order of the search tasks were randomly permuted.

For each task, the participants had to go through 4 stages:

- First, the participants were given a query and the corresponding background about the search task. They should read and memorize the search task carefully because it would not be shown again during the document reading.
- Second, one of the four documents was randomly chosen and presented to each participant. To avoid unnecessary distraction to the participants, the system only displayed the text content. An eye tracker was used to log participants' eye movements during reading in this stage.
- Third, the participants were asked to annotate the relevance of the given document in a four-level scale (4: very relevant; 3: fairly relevant; 2: somewhat relevant; 1: irrelevant).
- Finally, the document was presented to participants again and they were instructed to highlight relevant parts of text that were helpful for the search task. The participants were free to label any text (e.g., individual words, phrases, or whole sentences). If the document was totally irrelevant, they could skip this stage directly.

Table 2: The statistics of the data collected in the user study.

#Tasks	#Doc	#Participants	#Sesisons
15	60	29	435

In our study, we deployed a Tobii X2-30 eye tracker to capture participants' eye movements. Before the experiment, each participant should first go through a calibration process as required by the eye tracker. We used a laptop computer that had a 17-inch monitor with a resolution of 1600×900 and Google Chrome browser in our user study. To avoid potential distractions from text reading, all the experiments were conducted in the full-screen mode. The deviation of collected eye-tracking data is within the size of one word, , which enables us to use the eye-tracking data to estimate participants' reading attention at a word-level.

3.4 Statistics of Collected Data

Though the user study, we collected a dataset that consists of 435 relevance judgment sessions from 29 participants. The statistics of the collected data are shown in Table 2.

In the user study, each document was annotated by 7 or 8 participants. The inter-person agreement is measured by Cohen's κ. For the 4-level relevance annotation, the κ is 0.326, reaching a fair agreement level. If we convert the 4-level annotation into a binary relevance annotation by regarding 4: very relevant and 3: relevant as relevant and the rest as irrelevant, the κ for the converted binary relevance annotation reaches 0.757, suggesting a substantial agreement between the participants. A relatively high level of inter-person agreement indicates that participants indeed made reliable relevance judgments for the documents in the user study. According to the participants' annotations, among 60 sampled documents, 114 are 4: very relevant, 79 are 3: fairly relevant, 69 are 2: somewhat relevant, and 55 are 1: irrelevant.

We are also interested in whether the relevant text annotation (II. 4 in Figure 2) is reliable or not. Thus, in this study, we regard highlighting the relevant parts of text as making a binary annotation for each word in the document and further assume that these binary annotations are independent of each other. This enables us to use Cohen's κ to measure the inter-person agreement for the relevant text annotation task. The κ for relevant text annotation is 0.364, which suggests a fair agreement level among the participants.

Besides Cohen's κ, we also compute the precision of individual participants' relevant text annotations. We regard the parts of a document (i.e. a set of words in the document) that were highlighted by more than half of the participants who assessed it as the *true* relevant parts of the document. Then for the annotation of an individual participant, the precision of her annotation is given by the proportion of *true* relevant text in her highlighted text. The average precision of all the participants is 0.734, which shows that the relevant text annotation is fairly consistent across the participants.

4 TWO-STAGE READING MODEL

To address **RQ1** and **RQ2**, in this section we analyze the reading process during relevant judgement (i.e., eye movements and labeled relevant text) and propose a two-stage reading model. In addition, we compare participants' reading patterns on relevant and irrelevant document, respectively, to investigate how relevance affect the reading process.

Table 3: Reading behaviors on different vertical position in relevant and irrelevant documents. Independent t-test is performed to detect significant difference between relevant and irrelevant documents. Significant results are bold while * represents $p < 0.01$, two-tailed.

	Average Fixation Rate				
Position	0~20%	20%~40%	40%~60%	60%~80%	80%~100%
Relevant	0.238	0.253	0.230	0.224	0.160
Irrelevant	0.267	0.242	0.214	0.197	0.139
Diff	**-10.86%***	4.55%	**7.48%***	**13.71%***	**15.11%***
	Average Reading Time Per Word (msec)				
Position	0~20%	20%~40%	40%~60%	60%~80%	80%~100%
Relevant	106.3	112.61	92.81	82.93	52.47
Irrelevant	116.33	99.57	81.71	67.07	45.65
Diff	-8.62%	**13.10%***	**13.58%***	**23.65%***	**14.94%***

4.1 Preliminary Relevance Judgement

To investigate users' reading behavior in different parts of a document, we split the document content into five parts according to the vertical position. We first show the arrival time of each part in Figure 3. From the results, we can see that in general, the participants read the documents sequentially from top to bottom.

We then present the average fixation rate and average reading time per word for relevant and irrelevant documents in Table 3. The average fixation rate is the average likelihood that a word is fixated by the participant. Higher average fixation rate and longer average reading time per word indicate that the participants put more attention in this part of the document. First, we find that both fixation rate and reading time tend to decay as the vertical position increases, which indicates that users tend to read more carefully at the beginning. In terms of the difference between relevant and irrelevant documents, a significant difference (p-value < 0.01) in average fixation rate for the 0~20% part suggests that users paid more attention to the top 20% content on irrelevant documents. This is probably because that the participants were confused when reading the beginning of an irrelevant document and put more attention to confirm their irrelevant perception. For the lower part of the documents, the difference between these two document sets become apparent as the user tends to put much less attention to the lower part of irrelevant documents. This result indicates that the user has made a preliminary relevance judgement after reading the beginning of documents and this judgment will influence their follow-up reading behaviors. Therefore, the top 20%~40% content that draws more attention from users may be more important for the overall relevance judgment of the whole document.

4.2 Reading with Preliminary Relevance

We further investigate users' reading behavior in the documents with different perceived relevance. First, we compare users' fixation transition behavior on relevant and irrelevant documents. Then, we investigate how users' reading attention correlates with their explicit relevant text annotation.

4.2.1 Fixation Transition Behavior. Users' fixation transition behaviors can be split into three categories: *Forward*, *Regression*, and *Skim* [31]. The percentages of three transitions on the documents with different relevances are shown in Figure 4(a). To reduce the noise in eye-tracking data, we use a threshold of 200ms suggested by previous work [28] to filter out the fixations with a duration shorter than it. We also tried other thresholds, varying from 200ms to 500ms, and got similar results.

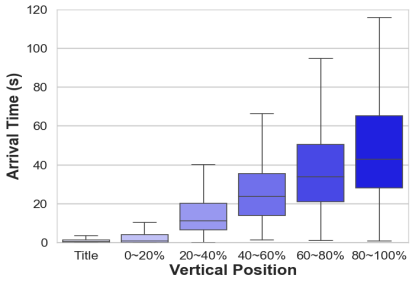
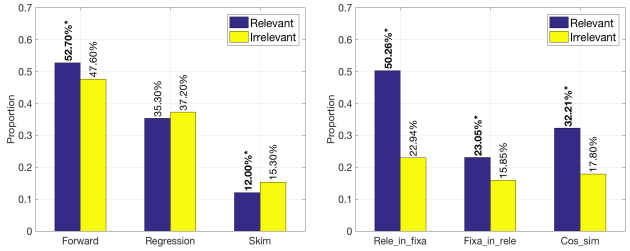


Figure 3: The arrival time in different vertical position.



(a) Transition behavior **(b) Correlations between reading text and relevant text**

Figure 4: Reading behaviors in the documents with different perceived relevances. Two-tailed t-test is performed to detect significant difference. Significant results are bold while * represents $p < 0.01$. "Rele_in_fixa" and "Fixa_in_rele" means the proportion of relevant text in the text that users fixate on and the proportion of the text that users fixate on in the relevant text, respectively. "Cos_sim" is cosine similarity.

It is observed that there are more forward transitions and less skim transitions on the relevant documents, with both differences significant with p-value < 0.01. This finding indicates that users tend to speed up their reading when they think the document is irrelevant and is similar to the Gwizdka et al.'s previous finding [18]. In addition, the percentage of regression on relevant documents is lower than that on irrelevant documents with a marginal significance (p-value = 0.011). It illustrates that as users reading faster in irrelevant documents, they are likely to go back to check some skipped content. We will further investigate what content users tend to skim in Section 5.2.

4.2.2 Reading Text vs. Relevant Text. Relevance judgment may depend on finding relevant text in the document. Therefore, we are also interested in how users read the relevant parts of text in the document, given different preliminary relevance judgments. We compared the users' reading text and their highlighted relevant text in the documents with different perceived relevances to investigate their relationships. Specifically, we measure the consistency between the

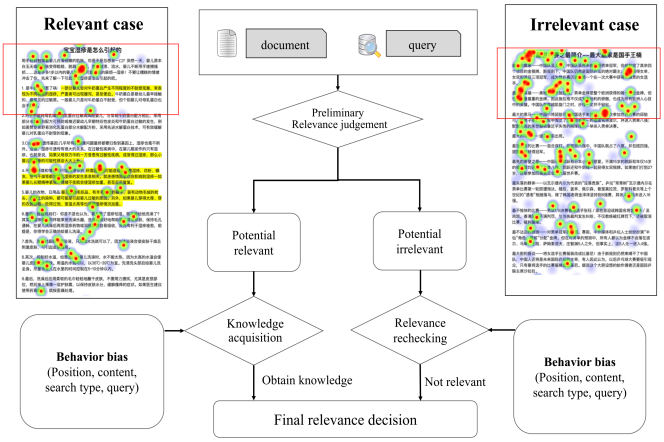


Figure 5: A two-stage reading model which contains a "Reading with Preliminary Relevance" stage (Stage 1) and "reading with preliminary relevance" stage (Stage 2).

eye-fixated reading text and labeled relevant text by their overlaps and the cosine similarity between them.

The results are shown in Figure 4(b). First, we can see that users' reading text is not always consistent with their labeled relevant text. Specifically, on relevant documents, users' reading text is more closely related to the relevant text in all three measurements, which indicates that users' eye-fixations are more reliable indicators for relevant text on the relevant documents. In addition, we find that for irrelevant documents, users still labeled some relevant text. When we asked the participants why they labeled relevant text on documents with lower relevance, most of them said that these documents only fulfill part of their information needs. Therefore, although there is some relevant text, they thought the whole document was not so relevant to the search task.

4.3 Two-stage Reading Model

Based on the analysis above, we believe that users' reading behavior during relevance judgement can be regarded as a two-stage process. Therefore, we propose a two-stage reading model, as shown in Figure 5.

When making relevance judgment for a document, users usually read the document sequentially from top to bottom, as shown in Figure 3. At the first stage, without any relevance perception, users exhibit similar behavior patterns when reading the beginning part of the document (about top 20% to 40%). However, their behaviors become significantly different on the later parts of relevant and irrelevant documents. This is due to that users have made a preliminary relevance judgment in the first stage and this judgment will influence their following reading behavior.

Then at the second stage, with the preliminary relevance judgement, users will have different reading behaviors on the documents with various relevance. When the document is perceived as irrelevant, the reading pattern is relatively disordered and can be associated with higher proportions of regression and skim. If the document is perceived as relevant, users will have a more sequential reading pattern and the reading text is more closely related to the labeled relevant text. These phenomena can be explained by that the preliminary relevance judgment affects users' intents. On a potentially relevant document, the user will try to explore and acquire more knowledge about the current topic or search task (Knowledge acquisition). On a seemingly irrelevant document, the user is still trying to find some relevance evidence so as to confirm or disprove the preliminary judgment (Relevance rechecking). Generally, the user will have a higher reading rate and put less attention to each word in the latter case.

Our model characterizes the reading pattern during relevance judgment. Existing attention models simply assume a uniform attention distribution over the whole document. However, in real scenarios, users tend to put more attention to the beginning parts of a document so as to make a preliminary relevance judgment for the document. Then in the second stage, user's attention distribution decays according to vertical position and is affected by the preliminary relevance judgement in the first stage. In addition, reading behaviors in these two stages are also influenced by other factors, such as position, lexical features, search task types, and query terms. We will further discuss it in the next section. Such analysis may help us better understand the cognitive process in relevance judgment task and guide the development more reasonable retrieval models.

4.4 Summary

In this section, we focus on **RQ1** and **RQ2**. For **RQ1**, we propose a two-stage reading model which contains a "preliminary relevance judgement" stage and "reading with preliminary relevance" stage. By modeling the reading behavior as a two-stage process, we characterize how the preliminary judgement in the first stage influences users' reading pattern in the second stage. For **RQ2**, we find that users' reading attention concentrates on the beginning part of document and decay gradually. This decaying tendency is more apparent for the irrelevant documents. In addition, in irrelevant documents, users' reading pattern is more disordered and contains more regressions and skims.

5 READING BEHAVIOR BIAS

In this section, we focus on investigating how different *behavior biases* influence users' reading process during relevance judgment. Specifically, we systematically study four factors that may alter users' reading behavior: position, lexical feature, search task type, and query terms.

5.1 Position Bias

Previous studies have investigated the position bias on the examination of the search results on SERPs [11, 17]. Higher-ranked results receive more user attention and have larger probabilities of being examined [25]. According to Table 3, users' fixation rate and reading time are both affected by the vertical position. To further demonstrate the effect of position bias on users' attention, we present the distribution of fixation numbers over different vertical positions in Figure 6.

We split the positions into 3 areas and find that users' reading attention decreases from Area ① to Area ③. Users tend to pay more attention to Area ① (top 30% of the document). This can be explained by that they need to first carefully read the beginning of the document to form a preliminary relevance judgment (Stage 1 of the two-stage reading model). The density of attention distribution

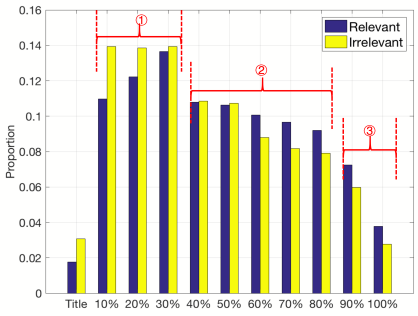


Figure 6: The proportion of fixation numbers in different vertical positions.

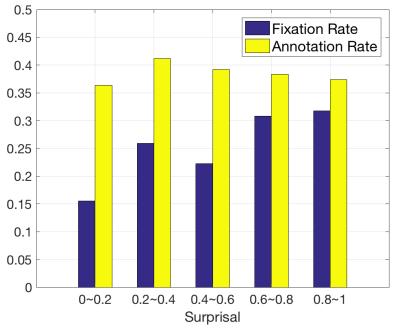


Figure 7: Fixation rate and relevant text annotation rate in different word surprisal.

in Area ② is significantly lower than that in Area ① (p-value < 0.01, two-tailed t-test), which indicates a transition to Stage 2, where users read the content faster. There is a sharp decrease in attention distribution in Area ③. It is probably caused by that some of the users already reach the final relevance judgment for the document and left after reading 80% of its content.

The above observations illustrate how the position bias affects users' reading behavior and attention distribution during relevance judgment. The results confirm that users' fixation attention is not uniformly distributed over the whole document, which implies that the content located at higher positions should be more important in determining the overall relevance of the document.

5.2 Lexical Feature

Psycholinguistic studies have shown that people read frequent words and phrases more quickly [23], thus we should also consider the influence of word complexity. We applied *surprisal*, which is the negative log-likelihood of a word in the context, to describe how unfamiliar a text is to users. By using the SRILM Toolkit [39], we built a bi-gram language model based on a large-scale online news data [40]. The relationship between word surprisal and attention is shown in Figure 7. The fixation rate and annotation rate is the likelihood that a word is fixated or labeled as relevant by the participant.

We can observe that the words with larger surprisal receive more reading attention but it is not obviously related to relevant text annotation. It is because surprisal is more closely related to low-level reading attention, which can be well captured by eye-movements [37]. On the other hand, relevant text annotation relies more on the semantic of text, thus surprisal is not a good indicator of relevant text.

5.3 Search Task Types

Users' cognitive processes vary in different information search tasks [24]. We follow Li and Belkin [24] and consider two product aspects of search tasks: *intellectual* and *factual*. In our study, we have 7 factual tasks and 8 intellectual tasks. The average proportion of relevant documents in two tasks is 60.4% and 61.3%, which is close to the proportion in the whole dataset. Users' reading behavior and relevant text annotation behavior for intellectual and factual tasks are compared in Table 4, where the values are normalized by the number of words.

Table 4: Reading behaviors and relevant text annotation rate in different search types. Two-tailed t-test is performed to detect significant difference. Significant results are bold while * represents $p < 0.01$.

	Fixation Rate	Reading Time	Annotation rate
Factual	0.232	96.14	0.222
Intellectual	0.201	75.07	0.299
Diff	**15.32%***	**28.07%***	**-25.76%***

It observed that users' fixation rate is lower and their reading time is shorter in intellectual tasks. However, the relevant text annotation rate in intellectual tasks is higher than that in factual tasks. This can be explained by that reading is more coherent in intellectual tasks, where users tend to skip the content that is already familiar for them. To validate this hypothesis, we compared a number of cases and found that there exists a number of relevant text that contains only a few (even no) fixation points in intellectual tasks. These texts are similar to the context and relatively understandable but indeed contribute to the search task. This indicates that there exists a gap between reading text and relevant text. The reading attention is influenced by many factors such as readability and coherence. Besides, this gap is more significant in intellectual tasks.

The above observation illustrates that the reading attention found by eye tracking devices is not consistent with the relevant text highlighted by the participants . Reading process is also controlled by other factors such as text readability, coherence, and users' preknowledge. The gap between reading text and relevant text is more significant in intellectual tasks compared to factual tasks.

5.4 Query Terms

Query is the explicit expression of users' information needs. Query-centric assumption proposed by Wu et al [41] strongly assumes that the relevant information for a query only locates in the contexts around query terms. Therefore, we try to inspect this assumption and investigate the influence of the contexts around query terms to users' reading attention. In addition, we also investigate this influence in different situations, i.e., in different vertical positions and in documents with different relevance. Specifically, stop words in query terms are removed to reduce the noise.

5.4.1 Impact of query-centric contexts. We aim to verify whether contexts around query terms have higher probability to draw users' reading attention and more relevant text annotation. Window

Table 5: Influence of query terms with different window size. Two-tailed t-test is performed to detect significant difference. Significant results are bold while * represents $p < 0.01$.

	Fixation rate				
win_size	0	1	3	5	10
Around query	0.279	0.246	0.239	0.236	0.231
Others	0.214	0.213	0.210	0.208	0.204
Diff	**30.14%***	**15.21%***	**13.78%***	**13.38%***	**13.12%***
	Annotation rate				
win_size	0	1	3	5	10
Around query	0.270	0.274	0.282	0.288	0.293
Others	0.258	0.256	0.251	0.243	0.226
Diff	4.80%	**7.07%***	**12.34%***	**18.53%***	**29.96%***

sizes from 0 to 10 are set to alter the width of context in words around the query terms. The influence of query terms to users' fixation rate and relevant text annotation rate is shown in Table 5.

We can observe that contexts around the query terms draw higher reading attention with p-value < 0.01, which indicates that query terms have strong impacts on users' reading process. When the window size is 0, it means that only the query terms are considered and they have drawn higher reading attention compared to other words. Typically, when the window size is larger than 0, the contexts around query terms have significantly higher probability to be annotated as relevant text. This indicates the query-centric assumption is reasonable, where the probabilities of reading attention and relevant text are both higher around the query terms. Therefore, by using this assumption, we can better find the location that draws users' reading attention and the relevant text in the relevance judgment task.

5.4.2 Position bias. As discussed in Section 5.1, position bias is shown to exist in users' reading process during relevance judgement. To better understand the influence of query terms, we eliminate the position bias by considering the above differences in different vertical positions. The window size is set as 5 and the result is shown in Table 6.

We can observe that at the top 40% position, users' reading attention of the contexts around query terms is similar to that of other words. However, in the rest content, query terms show significant impacts on users' reading attention. It verifies the existence of a preliminary relevance judgement stage in users' reading process. Since users tend to have higher reading attention and read more carefully in the beginning, there is no significant deviation of reading attention in the contexts around query terms. Later on, users tend to read faster since they have the pre-perceived relevance. Query terms become important in guiding users allocate their limited attention in this stage.

The above observation illustrates that query-centric bias may not affect users' reading attention in the beginning of a document because users consistently put more attention in this part, in spite of the relevance of the document. After the preliminary relevant judgement stage, contexts around query terms begin to show their impacts, which have higher probability to draw users' reading attention.

5.4.3 Impact of relevance. In the second stage of our proposed reading model, users' reading behavior is different with different pre-perceived relevances. To investigate the influence of query terms in different relevance-levels, we further conduct two-way

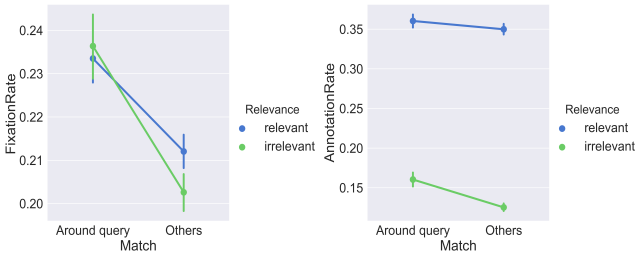
Table 6: Influence of query terms to reading attention at different vertical positions when window size is 5. Two-tailed t-test is performed to detect significant difference. Significant results are bold while * represents $p < 0.01$.

	Fixation rate				
Position	0~20%	20%~40%	40%~60%	60%~80%	80%~100%
Around query	0.255	0.256	0.244	0.241	0.168
Others	0.246	0.245	0.213	0.201	0.144
Diff	3.66%	4.49%	**14.55%***	**19.90%***	**16.67%***

ANOVA tests, that regard *Match* (whether the word is around query terms) and *Relevance* as factors, for both users' fixation rate and relevant text annotation rate. The result is shown in Figure 8.

For fixation rate, only the main effect of *Match* is significant, where $F(1,23808) = 111.56$, p-value < 0.01. In Figure 8(a), we observe that users' fixation rate is similar when focusing on contexts around the query terms. It illustrates that influence of contexts around the query terms to reading attention is independent of relevance.

For relevant text annotation rate, the main effect of *Match* ($F(1,23808) = 171.16$, p-value < 0.01) and *Relevance* ($F(1,23808) = 3555.18$, p-value < 0.01) are significant, and the interaction effect ($F(1,23808) = 9.89$, p-value < 0.01) is also significant. Figure 8(b) shows the average annotation rate under different relevances. Contexts around the query terms and higher relevance are associated with higher annotation rate. This indicates that query-centric assumption is more reliable to find relevant text in relevant document.



(a) Fixation Rate **(b) Annotation rate**

Figure 8: Two-way ANOVAs result of relevance influence in reading process.

5.5 Summary

In this section, we investigate the factors affecting users' reading process during relevance judgement, which is our research question **RQ3**. We found that position and lexical feature have more impacts on reading attention rather than labeled relevant text. Gap between users' reading attention and labeled relevant text is more significant in intellectual task. In addition, query-centric contexts indeed draw more reading attention but users do not rely too much on query in the beginning since they read more carefully for preliminary relevance judgement. Besides, the influence of query terms to reading attention is independent of relevance and they are more reliable to find relevant text in relevant documents.

Table 7: Features adopted in attention prediction.

Feature Category	Features
Structure	Position offset
	Vertical position
	Horizontal position
	Word length
Linguistic	TF-IDF
	Part-of-speech
	Word surprisal
	Word sparsity
	Sentence coherence
Query	Query exact match
	Query soft match

6 ATTENTION PREDICTION

In this section, we use different categories of features to predict users' fixation attention (fixation rate) and their labeled relevant text (annotation rate). In addition to the fine-grained word-level attention prediction, we also attempt to predict sentence-level and paragraph-level attention, which are defined as the average ratio of the fixated or highlighted words in a sentence or a paragraph. We regard the prediction as a regression problem and use Pearson's Correlation Coefficient (PCC) to evaluate the regression performance. The raw data of our experiments is available[1].

6.1 Features

The features we applied are listed in Table 7 and are categorized into three groups: Structure features, Linguistic features, and Query features. In addition to the factors discussed in Section 5, we also incorporate other text-based, static features into our prediction models. **Structure features** include the position and length of each word, which are commonly used in eye tracking studies [20]. **Linguistic features** are the text-based features that are independent of the query. Word sparsity is the word count in a large-scale online new corpus [40]. We also use this corpus to train word embeddings using Word2Vec algorithm. Sentence coherence features are then obtained by computing the minimum and average cosine similarity between the word embedding vectors in each sentence. **Query features** capture the similarity between the objective text and query terms. After removing the stop words in the queries, we use whether a word is in the query-centric contexts with window sizes of 1, 3, 5, 10 as the query exact match features and the cosine similarity between the word and the query as query soft matching features. Minimum, maximum, average, and sum values of all the word-level features are computed as the features for sentence-level and paragraph-level prediction.

6.2 Prediction Results

Attention prediction can be regarded as supervised regression problem. We perform a 5-fold cross validation to evaluate the performance of the regression model. We tried different regression models (e.g., Support Vector Machine (SVM), Gradient Boosting Regression Tree (GBRT), Logistic Regression (LR), Conditional Random Field (CRF) [20]) and found that their performances are similar. Therefore, we only show the results for GBRT in Table 8.

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

Table 8: Prediction results in different levels. (*/** indicate statistical significance at $p <$ 0.01/0.05 level compared to the best category, which is bold.)

	Category	Word	Sentence	Paragraph
Fixation Rate	Structure	**0.494**	**0.832**	**0.894**
	Linguistic	0.398*	0.790*	0.893*
	Query	0.216*	0.756**	0.803*
	All	0.505	0.836	0.902
Annotation Rate	Structure	**0.480**	0.454**	0.397
	Linguistic	0.458*	0.479	**0.471**
	Query	0.220*	**0.496**	0.429
	All	0.549	0.531	0.468

We first compare the performance of models based on Structural features, Linguistic features, and Query features, respectively. We can observe that for reading attention (fixation rate), structure features outperform the other two feature groups significantly and that query features are not so effective in predicting reading attention. This result is consistent with the findings in Section 5.1 that the position bias heavily influence user reading behaviors during relevance judgement, which suggests that we should incorporate the position bias into the computational attention models.

As for annotation rate, structure features also outperform other features at word-level. However, because the contexts around the query are more likely to be relevant, query features perform better than structure features at sentence-level. At paragraph-level, the semantic information is more important, therefore the model based on linguistic features has the best performance.

We also find that combining all the features consistently improve the performance of attention prediction. This finding further suggests that users' reading behavior and attention distribution during relevance judgment are affected by a variety of factors.

7 CONCLUSION

Human's reading behavior during relevance judgement is a complex cognitive process. In this paper, by conducting a carefully designed experiment, we found that users' reading process can be modeled as a two-stage process. First, in Stage 1, users tend to allocate a higher level of attention in reading the beginning (about 20% to 40%) of a document and make a preliminary relevance judgement. Then, in Stage 2, users take different reading strategies based on the preliminary judgement in the previous stage. They will gather evidence either to acquire knowledge or to validate the judgement. Detailed analysis of the experiment results further shows that users' reading process is affected by different factors, such as position bias, linguistic feature, search task types, and query terms. Specifically, we verified the *query-centric assumption* and discovered its subtle influence in different vertical position and relevance-level. Finally, we adopt GBRT to predict users' reading attention at different levels. Results show that query-independent features outperform query-dependent features, which indicates that users allocate attentions based on many signals other than query terms during relevance judgement.

Our proposed reading model can better explain users' cognitive process during relevance judgement and provides implications for improving the design of search engines. For example, we can use the

predicted attention distribution to improve existing retrieval models and summarization models. We can further infer the quality of relevance annotation by inspecting annotators' reading behavior.

In the future, we plan to study the reading behavior in a real search context with more complex page layouts and multi-modal elements such as images and videos. We are also interested in comparing the reading patterns of annotators and search engine users. We believe such studies can help us understand user's search process and provide insights for improving Web search engines.

8 ACKNOWLEDGEMENTS

This work is supported by Natural Science Foundation of China (Grant No. 61622208, 61732008, 61532011) and National Key Basic Research Program (2015CB358700).

REFERENCES

[1] Marco Allegretti, Yashar Moshfeghi, Maria Hadjigeorgieva, Frank E. Pollick, Joemon M. Jose, and Gabriella Pasi. 2015. When Relevance Judgement is Happening?:An EEG-based Study. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 719–722.

[2] Ioannis Arapakis, Konstantinos Athanasakos, and Joemon M. Jose. 2010. A comparison of general vs personalised affective models for the prediction of topical relevance. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 371–378.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. *international conference on learning representations* (2015).

[4] Peter Bailey, Peter Bailey, Peter Bailey, Peter Bailey, and Peter Bailey. 2014. Relevance and Effort: An Analysis of Document Utility. In *ACM International Conference on Conference on Information and Knowledge Management*. 91–100.

[5] Klinton Bicknell and Roger Levy. 2010. A rational model of eye movement control in reading. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1168–1178.

[6] Pia Borlund. 2003. The concept of relevance in IR. *Journal of the American Society for Information Science and Technology* 54, 10 (2003), 913–925.

[7] Hansjürgen Bucher and Peter Schumacher. 2006. The relevance of attention for selecting news content. An eye-tracking study on attention patterns in the reception of print and online media. *Communications* 31, 3 (2006), 347–368.

[8] Georg Buscher, Andreas Dengel, Ralf Biedert, and Ludger V. Elst. 2012. Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond. *Acm Transactions on Interactive Intelligent Systems* 1, 2 (2012), 9.

[9] Ye Chen, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. User Satisfaction Prediction with Mouse Movement Information in Heterogeneous Search Environment. *IEEE Transactions on Knowledge and Data Engineering* PP, 99 (2017), 2470–2483.

[10] Michael J. Cole, Jacek Gwizdka, Chang Liu, Ralf Bierig, Nicholas J. Belkin, and Xiangmin Zhang. 2011. Task and user effects on reading patterns in information search. *Interacting with Computers* 23, 4 (2011), 346–362.

[11] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *The ACM International Conference on Web Search and Data Mining*. 87–94.

[12] Edward Cutrell and Zhiwei Guan. 2007. What are you looking for?:an eye-tracking study of information usage in web search. In *Conference on Human Factors in Computing Systems, CHI 2007, San Jose, California, Usa, April 28 - May*. 407–416.

[13] D Drieghe, T Desmet, and M Brysbaert. 2007. How important are linguistic factors in word skipping during reading? *British Journal of Psychology* 98, Pt 1 (2007), 157.

[14] Denis Drieghe, Alexander Pollatsek, Barbara J. Juhasz, and Keith Rayner. 2010. Parafoveal processing during reading is reduced across a morphological boundary. *Cognition* 116, 1 (2010), 136–142.

[15] Ralf Engbert, André Longtin, and Reinhold Kliegl. 2002. A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision research* 42, 5 (2002), 621–636.

[16] Ralf Engbert, Antje Nuthmann, Eike M Richter, and Reinhold Kliegl. 2005. SWIFT: a dynamical model of saccade generation during reading. *Psychological review* 112, 4 (2005), 777.

[17] Laura A. Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking analysis of user behavior in WWW search. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 478–479.

[18] Jacek Gwizdka. 2014. Characterizing relevance with eye-tracking measures. In *Information Interaction in Context Symposium*. 58–67.

[19] Michael Hahn and Frank Keller. 2016. Modeling Human Reading with Neural Attention. In *Proceedings of the 54th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 85–95.

[20] Tadayoshi Hara, Daichi Mochihashi2 Yoshinobu Kano, and Akiko Aizawa. 2012. Predicting word fixations in text with a CRF model for capturing general reading strategies among readers. In *Proceedings of the 1st Workshop on Eye-tracking and Natural Language Processing*. 55–70.

[21] Peter Ingwersen and Kalervo Jrvelin. 2011. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer Publishing Company. 821–822 pages.

[22] B. J. Juhasz, S. J. White, S. P. Liversedge, and K Rayner. 2008. Eye movements and the use of parafoveal word length information in reading. *Journal of Experimental Psychology Human Perception and Performance* 34, 6 (2008), 1560.

[23] Marcel Adam Just and Patricia Ann Carpenter. 1987. *The psychology of reading and language comprehension.* Allyn & Bacon.

[24] Yuelin Li and Nicholas J. Belkin. 2008. *A faceted approach to conceptualizing tasks in information seeking*. Pergamon Press, Inc. 1822–1837 pages.

[25] Yiqun Liu, Chao Wang, Ke Zhou, Jianyun Nie, Min Zhang, and Shaoping Ma. 2014. From Skimming to Reading: A Two-stage Examination Model for Web Search. In *ACM International Conference on Conference on Information and Knowledge Management*. 849–858.

[26] Simon P Liversedge and John M Findlay. 2000. Saccadic eye movements and cognition. *Trends in Cognitive Sciences* 4, 1 (2000), 6–14.

[27] Tomasz D Loboda, Peter Brusilovsky, and Jöerg Brunstein. 2011. Inferring word relevance from eye-movements of readers. *Iui* 100, 1 (2011), 175–184.

[28] Lori Lorigo, Maya Haridasan, Hrnn Brynjarsdttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. 2008. Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the Association for Information Science and Technology* 59, 7 (2008), 10411052.

[29] Cheng Luo, Tetsuya Sakai, Yiqun Liu, Zhicheng Dou, Chenyan Xiong, and Jingfang Xu. 2017. Overview of the ntcir-13 we want web task. *Proc. NTCIR-13* (2017).

[30] Sandeep Mathias, Diptesh Kanojia, Kevin Patel, Samarth Agrawal, Abhijit Mishra, and Pushpak Bhattacharyya. 2018. Eyes are the Windows to the Soul: Predicting the Rating of Text Quality Using Gaze Behaviour.. In *Proceedings of the 56th annual meeting of the Association for Computational Linguistics*. To appear.

[31] 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.

[32] Yashar Moshfeghi and Joemon M. Jose. 2013. An effective implicit relevance feedback technique using affective, physiological and behavioural features. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 133–142.

[33] 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*. ACM, 257–266.

[34] K Rayner. 2009. Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology* 62, 8 (2009), 1457.

[35] Erik D Reichle, Alexander Pollatsek, Donald L Fisher, and Keith Rayner. 1998. Toward a model of eye movement control in reading. *Psychological review* 105, 1 (1998), 125.

[36] E. D. Reichle, K Rayner, and A Pollatsek. 2003. The E-Z reader model of eye-movement control in reading: comparisons to other models. *Behavioral and Brain Sciences* 26, 4 (2003), 445–476.

[37] Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition* 128, 3 (2013), 302–319.

[38] Alessandro Sordoni, Philip Bachman, Adam Trischler, and Yoshua Bengio. 2016. Iterative alternating neural attention for machine reading. *arXiv preprint arXiv:1606.02245* (2016).

[39] Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

[40] Canhui Wang, Min Zhang, Shaoping Ma, and Liyun Ru. 2008. Automatic online news issue construction in web environment. In *Proceedings of the 17th international conference on World Wide Web*. ACM, 457–466.

[41] Ho Chung Wu, Robert WP Luk, Kam-Fai Wong, and KL Kwok. 2007. A retrospective study of a hybrid document-context based retrieval model. *Information processing & management* 43, 5 (2007), 1308–1331.

[42] Xing Wu, Zhikang Du, and Yike Guo. 2018. A visual attention-based keyword extraction for document classification. *Multimedia Tools and Applications* (2018).

[43] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *international conference on machine learning* (2015), 2048–2057.

[44] Peng Zhang, Dawei Song, Yuexian Hou, Jun Wang, and Peter Bruza. 2010. Automata modeling for cognitive interference in users' relevance judgment. *Proc of Qi* (2010), 125–133.