Relevance Feedback with Brain Signals

ZIYI YE, XIAOHUI XIE, QINGYAO AI, YIQUN LIU, ZHIHONG WANG, WEIHANG SU, and MIN ZHANG, Quan Cheng Lab, DCST, Tsinghua University, Zhongguancun Lab, China

The Relevance Feedback (RF) process relies on accurate and real-time relevance estimation of feedback documents to improve retrieval performance. Since collecting explicit relevance annotations imposes an extra burden on the user, extensive studies have explored using pseudo-relevance signals and implicit feedback signals as substitutes. However, such signals are indirect indicators of relevance and suffer from complex search scenarios where user interactions are absent or biased.

Recently, the advances in portable and high-precision brain-computer interface (BCI) devices have shown the possibility to monitor user's brain activities during search process. Brain signals can directly reflect user's psychological responses to search results and thus it can act as additional and unbiased RF signals. To explore the effectiveness of brain signals in the context of RF, we propose a novel RF framework that combines BCI-based RF with pseudo-relevance signals and implicit signals to improve the performance of document re-ranking. The experimental results on the user study dataset show that incorporating brain signals leads to significant performance improvement in our RF framework. Besides, we observe that brain signals perform particularly well in several hard search scenarios, especially when implicit signals as feedback are missing or noisy. This reveals when and how to exploit brain signals in the context of RF.

CCS Concepts: • **Information systems → Information retrieval**; **Users and interactive retrieval**;

Additional Key Words and Phrases: Relevance feedback, brain computer interface, interactive information retrieval

1 INTRODUCTION

The application of **Relevance Feedback (RF)** is crucial to elicit additional information beyond the initial query, since queries submitted to the search engine are usually short, vague, and sometimes ambiguous [4]. RF [29, 69] method is widely applied to improve retrieval accuracy by acquiring feedback information in addition to the submitted query. In a standard RF process, users

should provide explicit relevance judgments on a given collection of documents. The search system then automatically re-ranks the search results with information extracted from the documents annotated as relevant or irrelevant by the user. RF paradigms effectively improve retrieval performance [29, 50, 59, 69], especially in the cases where users do not have a thorough idea of what information they are searching for or have difficulty transforming their search intents into queries [64].

Given that the standard RF process frequently places an additional cognitive burden on users, as it necessitates explicit relevance indications, prior research has explored a range of alternative signals to estimate the relevance of search results. These signals can be broadly categorized into two groups: pseudo-relevance signals and implicit signals. The basic idea of pseudo-relevance signals is simply treating the top-ranked documents as relevant and utilizing their terms for query rewriting, which is often applied to alleviate the problem of term mismatching [40, 43]. Pseudo-relevance signals do not necessarily require any user signals, making them easy to deploy and widely used. However, the quality of pseudo-relevance signals highly depends on the effectiveness of the initial retrieval [41] and is particularly unstable when the submitted query is vague or ambiguous.

Implicit signals (e.g., click and dwell time), on the other hand, are usually considered more reliable than pseudo-relevance signals, especially in interactive **Information Retrieval** (**IR**) scenarios [8, 59]. Nevertheless, implicit signals are indirect probes of relevance inferred from user's behaviors and thus are often biased and inaccurate [45]. While existing research attempts to mitigate these biases by utilizing additional signals such as eye-tracking and mouse movement, practical experience has revealed that these indirect signals still bring biases. For instance, the credibility of eye-tracking signals is limited as gazing at a document can not necessarily guarantee that its content is genuinely relevant [49]. Due to the inaccuracy and inheriting bias of pseudo-relevance/implicit signals, the effectiveness of RF is often limited [41]. Hence, the potential advancement of RF lies in acquiring more accurate and unbiased signals.

Recently, the emergence of neurological techniques has attracted researchers to explore interactive information systems with brain signals [19, 78]. As neurological devices become portable and affordable, it is possible to capture user's brain signals in realtime and build a practical search system with **brain–computer interface** (**BCI**) [12]. Beyond applying BCI to control search engines, existing studies [61, 79] have revealed underlying differences between brain responses to relevant and irrelevant search results. These differences indicate that brain signals can be acquired as novel alternative signals to measure relevance. In contrast to existing signals, brain signals directly reflect a user's psychological activities and, as a result, are less susceptible to the inherent biases associated with user behaviors. Yet, to the best of our knowledge, the potential benefits of brain signals in improving RF performance remain mostly unknown, especially when combined with existing signals.

In this article, we propose a novel RF framework combining brain signals with pseudo-relevance signals and click signals to improve RF performance. The proposed framework can be employed in two distinct RF settings, that is, **iterative RF** (**IRF**) and **retrospective RF** (**RRF**). IRF [1] dynamically re-ranks upcoming documents as more search results are presented to the user, which is preferable in situations where user signals are collected incrementally [8]. We are interested in analyzing how brain signals could improve the quality of RF while the search is ongoing. In contrast, the RRF setting focuses on re-estimating the relevance of historically examined documents after the search process ends. RRF cannot directly facilitate the ongoing search process, but it can help the potential search process with a similar search intent for other users [33]. In both IRF and RRF, our objective is to investigate whether brain signals can function as unbiased indicators of relevance. Additionally, we examine search scenarios where conventional user signals are biased, highlighting potential applications for BCI within IR.

To verify the effectiveness of IRF and RRF with brain signals, we conduct a lab-based Web search study and collect corresponding pseudo-relevance signals, click signals, and brain signals during the search process. Based on the collected dataset, we conduct document re-ranking tasks in the context of IRF and RRF, respectively. With extensive experiments, we demonstrate the effectiveness of the proposed framework, especially the benefits brought by brain signals. We observe that brain signals lead to an additional improvement of 8.8% and 7.4% in terms of NDCG@10 (Normalized Discounted Cumulative Gain) for IRF and RRF, respectively. Furthermore, we delve into search scenarios in which brain signals are more effective than pseudo-relevance and click signals. We observe that brain signals are particularly helpful in the cases where "bad click" happens, which may alleviate the "clickbait" issue (i.e., documents with misleading headlines attract bad click [46]). Besides, we have observed that brain signals can offer valuable relevance guidance for extracting information from non-clicks, a particularly advantageous feature for initiating RF before any click occurs. Drawing from these findings, we propose using a method that can boost RF performance by adaptively adjusting the combination weight of brain signals and other signals based on specific search scenarios.

In summary, our contributions are three-fold. (1) We devise a novel RF framework that combines pseudo-relevance signals, click signals, and brain signals to improve the performance of document re-ranking tasks. (2) We conduct a user study in Web search scenario to explore the improvement brought by brain signals in the proposed framework. (3) We investigate the possible search scenarios in which brain signals are more effective than existing signals. This reveals when and how we can bring the benefit of BCI into search systems.

The rest of this article is organized as follows. In the next section, we review related work in RF and neuroscientific approach. In Section 3, we introduce the Web search study and the data collection procedures. Then in Section 4, the preliminaries about the RF tasks (IRF and RRF) and the proposed RF framework are elaborated. Next, we present the experimental results and corresponding analyses to explore the impact of brain signals in the context of RF in Section 5. Finally, we conclude this work and discuss its applications and limitations in Section 6.

2 RELATED WORK

2.1 Relevance Feedback Signals

The standard RF process was based on explicit signals that would impose additional manual efforts and hurt the search experience. Hence, pseudo-relevance signals [43, 66] were often applied by simply assuming the top-ranked documents as relevant. However, Li et al. [41] observed that the performance of pseudo-relevance signals is often limited if the initial retrieval performance is weak. In addition to pseudo-relevance signals, implicit signals were also widely studied, in which a user's behaviors (e.g., click [69], dwell time [51], and eye-tracking [2]) are used to infer their preference. Among all the implicit signals, click signals were the most widely used and were thought to be an indication of relevance [69]. However, click signals are indirect probes of relevance [45] and may be biased in several search scenarios [23, 42, 71].

Traditionally, RF is based on only one type of signals (e.g., pseudo-relevance signals [43, 66, 81] or implicit signals [15, 68]), which limited their performance due to the inaccuracy and inheriting bias of existing RF signals [41]. To address this biased problem, prior literature has devised several robust architectures [11, 47, 80] to deal with biased signals. For example, Lv and Zhai [47] proposed to learn an adaptive coefficient that can avoid overvaluing RF information when RF signals may be unreliable. Besides, there existed a series of studies on click models [13, 84] have been conducted in an effort to estimate relevance from biased click signals.

Nevertheless, such approaches are still limited since user interactions are extremely varied and hard to probe [82]. Hence, potential advancement lies in acquiring more accurate RF signals. This

paper explores brain signals, which directly capture the user's thoughts and thus are unbiased and accurate. With empirical experiments, we verify that brain signals can bring additional improvement on top of existing signals, especially in situations where implicit signals are biased.

2.2 Relevance Feedback Techniques

Standard RF techniques required relevance assessments on a fixed batch of documents. Since pseudo-relevance signals are the most commonly used signals in standard RF, this kind of RF is also named as top-k RF. There were two main streams of top-k RF techniques, that is, **vector space model** (**VSM**)- based (e.g., Rocchio [63]) and **language model** (**LM**)-based (e.g., RM3 [40]). The VSM-based method refines the query vector to be closer to the relevant documents [63], while the LM-based method expands the query by selecting relevant terms [40]. Recently, with the developments of neural retrieval methods [58, 75], researchers have attempted to explore the effectiveness of RF with BERT-based ranker [43, 66, 67, 81]. For example, Wang et al. [66], Zheng et al. [85]adopted a VSM-based design to fuse BERT-based query embeddings with document embeddings. Zheng et al. [85] proposed BERT-QE which treats the top-ranked documents as additional queries to retrieve documents with a BERT-based re-ranker.

In addition to top-k RF, alternative RF techniques are devised in different settings. For example, Aalbersberg [1] propose IRF, which re-ranks upcoming documents after each user interaction in an incremental manner. This approach is particularly advantageous in practical search processes, especially when user signals are collected incrementally [8]. Besides transforming top-k RF techniques into IRF settings [1], recent works also devise novel IRF techniques for various scenarios, for example, conversational search [8] and product search [9]. In this paper, we first explore IRF as brain signals can provide real-time guidance for RF. In addition to up-coming documents, we also supplement the relevance re-estimation for historical documents, namely RRF in this article. RRF helps potential search processes with similar intents in the future by directly re-ranking the search results or constructing learning-to-rank models [33] with the estimated relevance signals. Conventionally, RRF relies on probing user behaviors (especially clicks) and strongly suffers from biased user behaviors [13, 33]. Hence, exploring alternative unbiased signals, that is, brain signals in RRF is valuable. The construction of IRF and RRF tasks are elaborated in Section 4.1.

2.3 Neuroscientific Approach for IR

There is increasing literature that adopts neuroscientific methods into IR scenarios [3, 12, 26, 28, 56]. These studies involved a variety of neuroimaging techniques, such as **functional magnetic resonance imaging** (**fMRI**) [54, 55], **magnetoencephalogram** (**MEG**) [34], and **electroencephalogram** (**EEG**) [19, 61, 79]. Amon]g these techniques, fMRI has the highest spatial resolution, which can help identify the topological distribution of cognitive functions during IR. For example, Moshfeghi et al. [55] examined the emergence of **Information Need** (**IN**) and identified activated brain regions. Their further study predicted the realization of an IN using fMRI voxels from a generalized set of brain regions (GM) and a unique set of regions for each individual (PM), respectively. Compared to fMRI, MEG, and EEG have higher temporal resolutions, which can help us understand the temporal dynamics of brain activity and construct real-time BCI applications. MEG produces better spatial resolution than EEG, but it is rather expensive and has strict requirements on the experimental environment where external magnetic signals should be shielded. On the other hand, EEG is more frequently used to construct real-time BCI applications due to its portability and affordability [35, 38]. For example, Chen et al. [12] propose an EEG-based demonstration search system that can obtain dictates in the **steady-state visual evoked potential** (**SSVEP**) for controlling search engines. Allegretti et al. [3] studied the time frame to distinguish human's brain response to relevant and irrelevant items. In their subsequent research [61], the graded

phenomenon of relevance was further investigated, that is, a search item could be partially relevant to the user.

Among these works, a common finding demonstrated by Allegretti et al. [3], Eugster et al. [26] is that brain signals can play as an indicator of relevance. For example, Eugster et al. [26] adopted brain signals to predict term relevance, and their further research utilized this prediction to interactively recommend other relevant information [25]. Allegretti et al. [3], Golenia et al. [27], on the other hand, predicts user's preference to image stimuli with brain signals. A more recent work Ye et al. [79] explored relevance in a more practical Web search scenario and observed that brain signals could detect the phenomenon of "good abandonment". However, their studies lacked applications for Web retrieval and limited user interactions (e.g., allowing interactions on **Search Engine Result Page (SERP)** to some degree.

What we add on top of existing work is that we explore whether brain signals can guide the RF process and improve document re-ranking. Recently, BCI applications have served as a new medium for human-system interaction in various real-life information systems, such as education [21], entertainment [20], virtual reality [65]. Therefore, understanding how BCI can enhance search performance, as it represents a typical channel for information acquisition, holds significant importance, particularly when considering research that indicates a BCI-enhanced search system is becoming realistic [12]. Our findings shed light on the specific scenarios and mechanisms through which BCI can elevate the search experience via RF.

3 DATA COLLECTION

To verify the effectiveness of brain signals and the proposed RF framework, we conducted a user study that simulates a practical search process. The participants were required to perform search tasks by interacting with Web pages, and their brain signals were collected during this process. The user study adheres to the ethical procedures approved by the ethics committee of the School of Psychology at Tsinghua University. In compliance with established ethical guidelines, we have taken multiple measures to safeguard the privacy of the participants, including anonymizing the collected data, obtaining informed consent from participants before the study, and allowing the participants to interrupt the experiment at any time.

3.1 Participants

We recruited 21 participants (8 females and 13 males) aged from 19 to 27 (M[1] = 23.85, SD[2] = 2.28) from a Chinese public university. The amount of participants is analogous to prior EEG-based studies (e.g., 18 in [79] and 23 in [61]) and the estimated sample size for the feature analysis in Section 5.1 is 18 (statistical power = 0.8, α = 0.05). All participants are native Chinese speakers and self-reported that they are familiar with the usage of search engines. The experimental procedures encompass a one-hour neurological experiment, resting periods totaling thirty minutes, a thirty-minute session for hair shampooing (before and after the main experiment), a thirty-minute duration for questionnaires, experimental instructions, and wearing the EEG cap. The participants were compensated 15.0$ per hour, amounting to a total payment of approximately 37.5–45.0 dollars.

3.2 Stimuli Preparation

For the lab study, we constructed a dataset containing 100 queries and their corresponding documents (an average of 39.5 documents for each query). The dataset is publicly available.[3]

---

[1]Mean value.

[2]Standard deviation.

[3]https://github.com/THUIR/Brain-Relevance-Feedback

Table 1. Examples of the user Study Queries (Translated into English)

Query	Source	Task descriptions
prophet	iMine 0001	(i) find information about "prophet" in Islam; (ii) explore the concepts of "prophet" in the general domain; (iii) search for a movie named "prophet" with high-quality audio source.
Persian cat	iMine 0002	(i) learn about the concept and characteristics of Persian cat;(ii) download pictures of Persian cat; (iii) learn about the market price of Persian cat;(iV) read books with Persian cat themes.
multiplication tables	iMine 0042	(i) download multiplication tables; (ii) learn the tips for memorizing multiplication tables.
pen	TREC 20582	(i) learn about fountain pens and their origins; (ii) check the brand and price of fountain pen.
tea	TREC 20906	(i) learn about the benefits of tea, such as nutritional components and efficacy; (ii) learn about the varieties of tea for a course presentation; (iii) explore the preparation methods of tea; (iv) exploring different commercial brands of tea.
milk experiments	TREC 21092	(i) learn information related to chemical reactions between litmus and milk; (ii) learn information related to sensory experiment on milk or yogurt.; (iii) learn about general science experiments related to milk.

Query and document set construction. We selected queries from NTCIR-11 IMine [44] and TREC-2009 Million Query Track [14]. We chose these datasets for the following reasons: (1) The datasets are independent of one another. (2) Most of their queries are short and on a broad topic, which means that the potential benefits of RF could be easily observable. Hundred candidate queries are selected from these datasets, 50 from NTCIR-11 IMine and 50 from TREC-2009 Million Query Track. For NTCIR-11 IMine, all Chinese queries (a total of 50) are included. For TREC-2009 Million Query Track, we selected 50 queries with familiar topics among Chinese users (e.g., "tea", and "pen"), and translated them into Chinese. However, we did not include queries with specific entities that may not be familiar to all Chinese users, for example, "Kansas City", "Vonage", and "Chicago Defender". Table 1 displays sample queries, while the complete query dataset is publicly accessible.[3]

After that, we utilized a popular Chinese search engine Sogou[4] to retrieve the corresponding documents for each query. Note that we can not reuse Sogou's datasets (e.g., SRR [83]) since they only contain plain text of documents. This limitation extends to the absence of images in landing pages and SERP, which are required to construct the user study system. To address this gap, we crawl the top five SERPs by submitting these queries to Sogou. Then, for each document, we extracted its snippet (document abstraction in the SERP) and crawled its landing page (the Web page that a user is directed to after clicking the document). The crawling process proceeded with the Python package "selenium". All web pages were manually checked, and we dropped documents whose landing pages were invalid or failed in crawling. This process generates an average of 39.5 documents for each query, and each document consists of snippet content and landing page content.

---

[4]www.sogou.com.

Task description construction. As most of the queries are short and broad, we generated 2–5 possible task descriptions for different subtopics regarding the query, as shown in Table 1. For example, for the query "prophet", the task descriptions include: (i) find information about "prophet" in Islam, (ii) explore the concepts of "prophet" in the general domain, and (iii) search for a movie named "prophet" with high-quality audio source. To generate the task descriptions, we adopted a similar procedure as Liu et al. [44] proposed in the subtopic mining subtask. Two Ph.D. students majoring in IR are recruited as domain experts to generate the subtopics for each query with the following steps: first, they independently clustered the documents into at most five groups and generated a task description for each group according to the common subtopic of the documents. Second, they were required to compare their clustering documents and the corresponding task descriptions. After the comparison, similar task descriptions were preserved and merged, while some dissident task descriptions were discarded. Finally, they discussed and modified each task description to reach an agreement. As a result, the construction process averagely generated 2.6 (SD^2 = 1.0) task descriptions for each query.

Snippet Annotation. To evaluate the RF performance, we recruited nine additional annotators with experience in search engine usage from a Chinese public university. They are required to judge the relevance of each document's snippet regarding the task descriptions given by the experts. The judgment is binary, that is, relevant or irrelevant. Each snippet was judged by three annotators, and the majority vote decided whether it was relevant. The Fleiss's κ of relevance judgment from the annotators is 0.76, reaching a substantial agreement. Note that we did not judge the relevance of the landing pages in advance due to the prohibitive cost. Instead, the participants were required to judge the relevance of their examined landing pages after the search process.

3.3 Experimental Procedure

Overall pipeline. The user study began with a questionnaire to collect demographic information (e.g., age and major) and search habits (e.g., how often do you use search engines?). Then, participants were instructed on the main task procedure and were informed that they could terminate the experiment and receive payment according to the time duration (13$ per hour). Next, participants were required to finish a training process, which included two search tasks to get familiar with the search process. After that, the main task began, in which the participants were supposed to accomplish as many search tasks as possible in 1.5 hours. During the main task, they were allowed to rest after every five tasks. Averagely, one participant will accomplish 36.8 (SD^2 = 10.3) search tasks in one hour. Finally, the participants need to fill in a post-questionnaire about their search experience.

Main task procedure. Figure 1 delineates the sequential procedures, spanning S_1 to S_4, of a main task. The interface of the user study system is expounded in Section 3.4 and visually represented in Figure 2. The system can be deployed using our publicly available code.^{3}

(S_1) The participants view a query and one of its corresponding task descriptions. To mitigate the incidental effect related to the search task, we randomized the selection process of the queries, the task descriptions, and the displaying sequence of documents. Examples of task descriptions are shown in Table 1.

(S_2) A fixation cross is presented for 0.5 seconds to indicate the location of the forthcoming stimulus. Then a document's snippet extracted from the SERP (a snippet includes a title, an abstract, and an optional image, as shown in Figure 1) is shown to the participant. The document's snippet is set unclickable in the first 2 seconds, and then it will be wrapped by a highlighted border, indicating that it becomes clickable. This procedure, following existing works [55, 79], ensures that brain activity related to the motor response of clicking would not be contained during the first 2 seconds. The screenshots of the snippet page are presented in Figure 2(c) and Figure 2(d), respec-
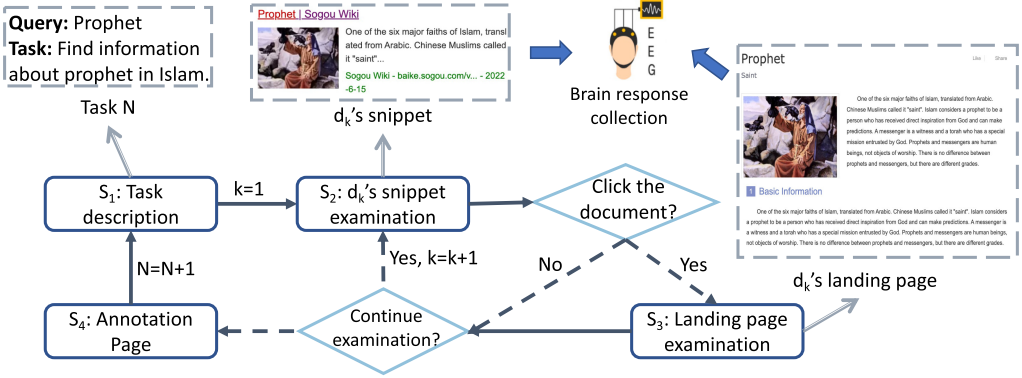
Fig. 1. The structure of the main task exemplified by the query "Prophet" (iMine 0001). (zoom in for details of example snippet and landing page).

tively. Note that there is some strictness in this step to ensure the presented information aligns with the participant's perception. Similar restrictions also apply to previous studies [5, 19, 26]. If the participant clicks the document, they will enter the document's landing page (jump to S_3). Otherwise, they can continue the examination (stay in S_2 with the next document's snippet presented) or end the search process (jump to S_4). To be consistent with the realistic search process, we allow the participants to revisit or click prior documents at any time in S_2. However, we only collect brain responses to each document during the first examination for further analysis.

(S_3) The landing page crawled from the corresponding document's link is presented, which generally contains more specific information than the snippet on the SERP (as shown in Figure 1). Similar to S_2, the examination will last for at least 2 seconds, and the participant can choose to continue the examination (jump to S_2 with the next document's snippet presented) or end the search (jump to S_4).

(S_4) The documents examined by the participant will reappear to collect the participant's relevance judgment. The judgment is conducted with a four-point Likert scale (ranging from 1 ("totally irrelevant") to 4 ("perfectly relevant"). The participants are required to annotate the relevance of the landing pages (in S_3) and the snippets (in S_4) independently since sometimes the relevance of a document's snippet may not be aligned with its landing page. For instance, there may be occasions where the snippet appears enticing, yet the corresponding landing page fails to meet the user's actual needs [46]. For example, when the snippet relevance is 4, the landing page relevance of the same document can be judged as 1 if the landing page does not satisfy the participant (i.e., "bad click" happens).

3.4 Interface of the Lab Study

Figure 2 showcases the interface of the Web search lab study using an example query "tea" (TREC 20582). On all pages, the background color is set to black, a common practice in neurological experiments to minimize interference from irrelevant factors [57].

In Figure 2(a), the login page is displayed, where participants enter their student ID (or a randomly generated ID if they are not students) along with a randomly assigned user ID. Post login, search tasks are sequentially presented to the participants. To complete each search task, participants undertake four steps, elaborated in Section 3 and Figure 1.

Figures 2(b)–2(f) provide illustrations and screenshots for steps S_1-S_4 (as detailed in Section 3). Figure 2(b) portrays the search task description page where participants are instructed to thor-
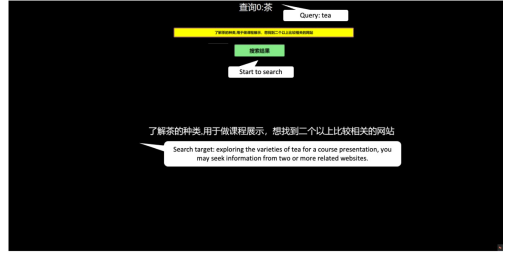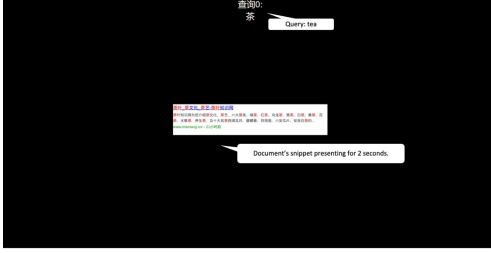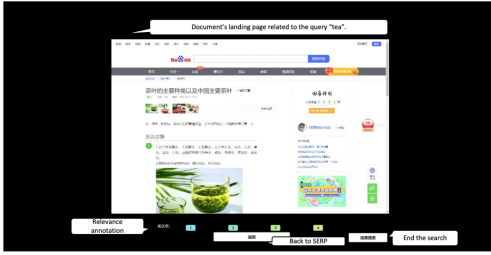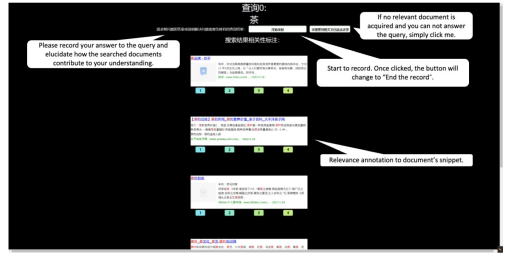
(a) The login page.

(b) S_1: The search task description page.

(c) S_2: The snippet page (within first 2 seconds) .

(d) S_2: The snippet page (after 2 seconds) .

(e) S_3: The landing page.

(f) S_4: The annotation page.

Fig. 2. Interface of the lab study. Figure 2(a) depicts the login page. Figure 2(b) illustrates the search task description page corresponding to step S_1 in the main task (detailed in Section 3). Meanwhile, Figures 2(c) and 2(d) correspond to step S_2, Figure 2(e) corresponds to step S_3, and Figure 2(f) corresponds to step S_4.

oughly read the query term (e.g., "tea") at the top, followed by the task descriptions (e.g., "exploring the varieties of tea for a course presentation, seeking information from two or more related websites"). Subsequently, participants can click "start to search" to proceed to step S_2. Figures 2(c) and 2(d) depict the document snippets page in step S_3. Initially, for the first two seconds, the document's snippet is rendered unclickable (as demonstrated in Figure 2(c)) to ensure the exclusion of brain activity related to the motor response of clicking [55, 79]. Post this period, participants are provided with four optional actions, illustrated in Figure 2(d): (1) click the document to access the corresponding landing page, (2) click "Next document" to examine the following document, (3) click "End the search" to proceed to the annotation page, (4) examine and interact (e.g., click) with previous documents displayed on the right side of the screen. Figure 2(e) demonstrates the landing page interface where participants can scroll to inspect all content on the landing page. Upon examination, participants may annotate a relevance score for the landing page and return to S_2 by clicking "Back to SERP". If participants end the search in S_2 or S_3, they are directed to the annotation page as illustrated in Figure 2(f). On this page, participants are supposed to verbalize a concise answer to the search tasks, following which they are required to annotate each document snippet with a relevance score.
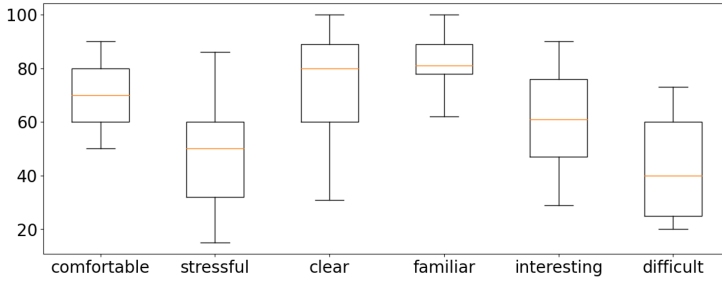
Fig. 3. Box plot of participants' evaluations on their search experience. The orange diamond represents the mean value.

3.5 Apparatus and EEG Preprocessing

The user study proceeded with a desktop computer (monitor size: 27-inch, resolution: 2,560 × 1,440) and the Google Chrome browser. A Scan NuAmps Express system (Compumedics Ltd., VIC, Australia) and a 64-channel Quik-Cap (Compumedical NeuroScan) are deployed based on the International 10âĂŞ20 system to capture the participant's EEG data (electrical activity of the brain). The impedance of the channels was calibrated under 25 $k\Omega$ and the sampling rate was set at 1,000 Hz.

To standardize the preprocessing process, we segmented 2,000 ms of EEG data upon the stimulus (i.e., snippet or landing page) presented to the participant. After that, we processed the EEG data with the following steps: re-referencing to averaged mastoids, baseline correlation with pre-stimulus periods (500 ms), low-pass of 50 Hz and high-pass of 0.5 Hz filtering, and down-sampling to 500 Hz. The preprocessing process is analogous to prior literature [61, 79] excepting that time-wasting artifacts removal techniques such as parametric noise covariance model [30] and **Independent Component Analysis** (**ICA**) [70] were not adopted. This is because we aim to assess the possibility of building a real-time RF method with brain signals while maintaining acceptable preprocessing time.

3.6 Questionnaire Analyses

We investigate the participant's user study experience with a post-questionnaire. We ask the participants to report their perceptual difference between the user study procedure and their daily search procedure using a five-level Likert scale (very small, small, neither small nor big, big, and very big). Most participants (57.1%) feel the difference is small, followed by neither small nor big (23.8%), big (14.3%), and very small (4.8%). This indicates that our user study design is close to reality. We also asked the participants about the major differences they felt, of which not allowing query reformulation is reported most often. As the first step to explore RF with brain signals, we limited the user study to an ad-hoc search scenario and left the study of multi-turn retrieval as future work.

Besides, we collect participants' evaluations on their search experience by asking "The search tasks in the user study is [comfortable/stressful/clear/familiar/interesting/difficult]" with a 100-grade scale where 100 indicates "strongly agree" and 0 indicates "strongly disagree". Figure 3 presents the box plot of the participants' evaluation gathered from the questionnaires. These results indicate that participants found the tasks comfortable, clear, familiar, and interesting but neither stressful nor difficult.

3.7 Statistics of the Collected Data

The dataset consists of 979 search tasks collected from 21 participants. The participants averagely accomplished 46.6 (SD[2] = 16.6) search tasks. Within a search task, they averagely examined 10.9

Table 2. Notations of the Relevance Feedback Framework

Notation	Definition
$q, \mathcal{D}$	the query, the documents regarding the query
$\mathcal{D}_h, \mathcal{D}_u$	the historical documents, the up-coming documents
$\mathcal{R}^p, \mathcal{R}^c$	$\mathcal{D}_h$'s pseudo-relevance scores, click-based relevance scores
$\mathcal{X}^s, \mathcal{X}^l,$	brain responses to $\mathcal{D}_h$'s snippet, landing page
$\mathcal{R}^s, \mathcal{R}^l$	$\mathcal{D}_h$'s brain-based relevance scores for its snippet, landing page
$\mathcal{R}^{bs}$	$\mathcal{D}_h$'s brain-based relevance scores for the document
$\mathcal{R}^{it,h}, \mathcal{R}^{re}$	$\mathcal{D}_h$'s combined relevance scores in IRF, RRF
$\mathcal{R}^{it}$	$\mathcal{D}_u$'s estimated relevance scores in IRF
θ^{it}, θ^{re}	combination parameters in IRF, RRF
$\mathcal{R}^{gu}, \mathcal{R}^{gh}$	the ground truth relevance scores of $\mathcal{D}_u, \mathcal{D}_h$

documents (10,670 documents in total) and clicked 1.9 of them (1,820 clicks in total). According to the participants' annotations, the average relevance of the document's snippets is 1.72, while that of landing pages is 2.67 (note that only clicked documents get landing page annotations). The annotations made by the participants for the 10,670 document snippets show a correlation with the third-party annotation at Kendall's τ of 0.66 ($p < 1e^{-3}$). This demonstrates that there is a high level of consistency between participants and third-party annotators in their judgment of the relevance of a document. Besides, the datasets contain the participants' brain responses, amounting to 10,670 and 1,820 EEG segments collected during their examination on snippets and landing pages, respectively.

4 METHOD

This section elaborates the RF problem definition, the steps of the proposed RF framework (as visualized in Figure 4), and the training and evaluation pipeline of RF. Table 1 presents the notations used within the RF framework.

4.1 Problem Definition

We assume a user issues a query q, and the documents list associated with the query q is $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$. Suppose the user examined h_{max} documents under the query q before they end the search. At a certain state during the process, h (from 1 to h_{max}) historical documents $\mathcal{D}_h = \{d_1, d_2, \ldots, d_h\}$ have already been examined, and there still exist $n-h$ up-coming documents $\mathcal{D}_u = \{d_{h+1}, d_{h+2}, \ldots, d_n\}$ have not been examined.

With the feedback signals acquired from the interactions on historical documents $\mathcal{D}_h$, the proposed framework aims to facilitate two RF tasks, that is, IRF and RRF. IRF re-ranks upcoming documents $\mathcal{D}_u$ by re-estimating their relevance $R^{it} = \{r^{it}_{h+1}, \ldots, r^{it}_n\}$. It can be adopted in real-time search scenarios: as the number of examined documents h increases, IRF iteratively re-ranks $\mathcal{D}_u$ to benefit the current search process. On the other hand, RRF re-ranks historical documents $\mathcal{D}_h$ by re-estimating their relevance $R^{re} = \{r^{re}_1, \ldots, r^{re}_h\}$. It does not aim at re-ranking documents for the current search process, but it can be utilized to provide a better ranking list in prospective search processes with similar intents.

To evaluate IRF and RRF, we apply ranking-based metric Π (e.g., **Normalized Discounted Cumulative Gain** (**NDCG**) [32]). Suppose the ground truth relevance for $\mathcal{D}_u$ and $\mathcal{D}_h$ are $R^{gu} = \{r^{gu}_{h+1}, \ldots, r^{gu}_n\}$ and $R^{gh} = \{r^{gh}_1, \ldots, r^{gh}_h\}$, respectively, where r^{gu}_i ($i > h$) or r^{gh}_i ($i \leq h$) is the ground truth relevance of the ith document. The performance of IRF and RRF can be measured as $\Pi(R^{gu}, R^{it})$ and $\Pi(R^{gh}, R^{re})$, respectively.
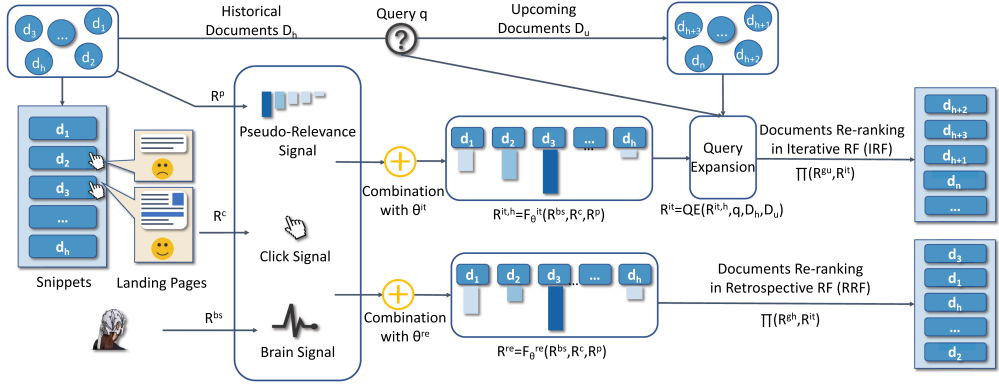
Fig. 4. During the search process, the user's brain signals and click signals are collected in addition to the pseudo-relevance signals provided by the system. These signals are then combined into two relevance scores for IRF and RRF tasks, respectively. Then the combined relevance score for RRF is directly applied to re-estimating the relevance of the historical documents. Besides, the upcoming documents can be re-ranked with a query expansion module which extracts useful contents in historical documents with the combined relevance score for IRF.

4.2 RF Framework

The proposed RF framework includes the following steps: (1) Pseudo-relevance signals, click signals and brain signals are independently transformed into base relevance scores R_p, R_c, and R_{bs}, respectively. The brain-based relevance score R_{bs} is chosen from the brain-based relevance scores of the document's snippet R^s and the document's landing page R^l. (2) For the ith historical document $d_i \in \mathcal{D}_h$, the feedback signals $r_i^p \in R_p$, $r_i^c \in R_c$, and $r_i^{bs} \in R_{bs}$, are combined into relevance scores $r_i^{it,h} \in R^{it,h}$ and $r_i^{re} \in R^{re}$ for IRF and RRF, respectively. For simplicity, we utilize F_θ to denote the combination using parameter θ, e.g., $R^{it} = F_{\theta^{it}}(R_{bs}, R_c, R_p)$. (3) In RRF, the historical documents $\mathcal{D}_h$ are directly re-ranked according to the combined relevance score R^{re}. (4) In IRF, we generate the relevance score for upcoming documents by a query expansion module QE, which can be formulated as $R^{it} = QE(R^{it,h}, q, \mathcal{D}_h, \mathcal{D}_u)$, where the combined relevance score for historical documents $R^{it,h}$ are applied to balance the importance of each document for query expansion module QE.

4.2.1 Signals Preparation. To generate high-quality RF, we acquire and combine base relevance scores independently estimated from pseudo-relevance signals, click signals, and brain signals, which are detailed as follows:

Pseudo-relevance score. By measuring the semantic similarity between q and each document $d_i \in \mathcal{D}_h$, we generate the pseudo-relevance scores $R_p = \{r_1^p, r_2^p, \ldots, r_h^p\}$, where $r_i^p \in [0, 1]$ is the ranking score of d_i. For a query q and a document d, we adopt a BERT re-ranker [58, 77] to measure their semantic similarity, denoted as $BERT(q, d)$. The BERT re-ranker is initialized by fine-tuning BERT-Chinese[5] on $T^2Ranking$ dataset[6] with the same procedures and available codes in the dataset's original paper [72]. The dataset contains human annotations on queries and documents extracted from Sogou's search log, which is similar to the dataset constructed for our user study. Besides, we use the snippet content to represent the document d's for estimating the semantic

---

[5]https://github.com/ymcui/Chinese-BERT-wwm
[6]https://github.com/THUIR/T2Ranking

similarity, as it reflects what users typically observe on SERPs. This choice effectively captures the primary theme of the landing page while filtering out extraneous or irrelevant details.

Click-based relevance score. The click-based relevance score $R^c = \{r_1^c, r_2^c, \ldots, r_h^c\}$ is generated from user's clicks behaviors, where $r_i^c = 0$ (or 1) indicates the user abandons (or clicks) the ith document.

Brain-based relevance score. With EEG devices, the user's brain responses to snippet content and landing page content are collected as $\mathcal{X}^s = \{x_1^s, x_2^s, \ldots, x_h^s\}$ and $\mathcal{X}^l = \{x_1^l, x_2^l, \ldots, x_h^l\}$, respectively, where x_i^s and x_i^l are the brain responses to the ith document's snippet and landing page, respectively. A sample of the brain signals, denoted as $x \in \{x_i^s, x_i^l\}$, can be presented as a vector in the space $\mathbb{R}^t$, with t denoting the length of EEG features. With a brain decoding model G (elaborated in Section 5.1), X^s and X^l are then transformed into brain-based relevance $R^s = \{r_1^s, r_2^s, \ldots, r_h^s\}$ and $R^l = \{r_1^l, r_2^l, \ldots, r_h^l\}$, respectively, where r_i^s and r_i^l are real numbers in $[0, 1]$ indicating the brain-based relevance score for the snippet and landing page, respectively. As the user's brain responses to the landing page content, represented by x_i^l, cannot be acquired when they do not click the document and enter the landing page, the corresponding value of r_i^l is masked and considered as unavailable in such situations.

Based on R^l and R^s, we generate $R^{bs} = \{r_1^{bs}, \ldots, r_h^{bs}\}$, which presents the brain-based relevance scores of the documents $\mathcal{D}_h$. We use different principles to generate R^{bs} for IRF and RRF, respectively. The reasons and details are as follows:

(i) In IRF, we extract information from $\mathcal{D}_h$ in the query expansion procedure. As the snippet usually presents the document more briefly and contains less noise than the landing page, we utilize the snippet content rather than the landing page content to represent the document. Hence, we can simply assign $r_i^{bs} = r_i^s$.

(ii) In RRF, we aim to re-rank the documents $\mathcal{D}_h$ themselves. If a document d_i has an attractive snippet but a landing page of low quality, it often leads to a poor user experience [46] and should not be evaluated as a satisfying document. Therefore, if r_i^l is available (not a masked value), r_i^{bs} is assigned the value of r_i^l; otherwise, we assign r_i^{bs} as r_i^s.

4.2.2 Relevance Score Combination. For the ith document, its combined relevance scores r_i^{it} and r_i^{re} are combined with various relevance scores from pseudo-relevance signals (r_i^p), click signals (r_i^c), and brain signals (r_i^{bs}), which can be formulated as:

$$r_i^{it,h} = \theta^{it,bs} \cdot r_i^{bs} + \theta^{it,c} \cdot r_i^c + \theta^{it,p} \cdot r_i^p \qquad (1)$$

$$r_i^{re} = \theta^{re,bs} \cdot r_i^{bs} + \theta^{re,c} \cdot r_i^c + \theta^{re,p} \cdot r_i^t, \qquad (2)$$

where $\Theta^{it} = \{\theta^{it,bs}, \theta^{it,c}, \theta^{it,p}\}$ and $\Theta^{re} = \{\theta^{re,bs}, \theta^{re,c}, \theta^{re,p}\}$ are the combination parameters. It is critical to understand that we use different combination parameters for IRF and RRF for the following reasons:

(i) We utilize different principles to generate R^{bs} for IRF and R^{bs} for RRF, hence the combination weights should be different.

(ii) $R^{it,h}$ is applied as an indication to extract text information from $\mathcal{D}_h$'s snippet while R^{re} is directly applied to re-estimate the relevance of $\mathcal{D}_h$. Hence inherent differences exist between them, for example, $R^{it,h}$ emphasizes whether the document's snippet content is relevant while R^{re} should also reflect the document's quality beyond the text.

Besides, it is also important to note that $R^{it,h}$ and R^{re} can be either fixed or adjustable to different search scenarios, which are detailed in Section 5.2.1 and Section 5.4, respectively.
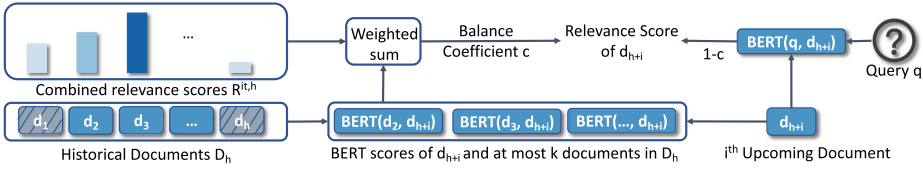
Fig. 5. The query expansion module.

4.2.3 Query Expansion. This section elaborates on the query expansion module QE, as depicted in Figure 5. The query expansion module utilizes the combined relevance score $R^{it,h}$ and the historical documents $\mathcal{D}_h$ to re-rank upcoming documents $\mathcal{D}_u$. As BERT has become a prevalent base ranker for RF [43, 81, 85], we adopt a BERT-based query expansion module into our RF framework with the following extensions: (i) We extend this module into an interactive setting for IRF. (ii) We bring in the estimated relevance score combined from various signals. The details of the query expansion module are as follows:

First, we select a maximum of k (a fixed value set as 10) documents from the historical documents $\mathcal{D}_h$, denoted as $\mathcal{D}_s = \{d_{s^1}, \ldots, d_{s^k}\}$, based on their combined relevance scores $r_i^{it,h}$ ($i \in 1, 2, \ldots, h$), which surpass those of other documents. Second, for each upcoming document d_{h+i}, we calculate the BERT score $BERT(d_s, d_{h+i})$ for each document d_s in $\mathcal{D}_s$. Third, the calculated BERT scores are weighted sum based on the combined relevance score $R^{it,h}$ to acquire the relevance score r_{h+i}^f for the $(h+i)^{th}$ document:

$$r_{h+i}^f = \sum_{j=1}^{k} \frac{e^{r_{s_k}^{it,h}}}{\sum_{l=1}^{k} e^{r_{it,h}^{s_l}}} \cdot BERT(d_{s_j}, d_{h+i}). \tag{3}$$

Fourth, the relevance score for the ith upcoming document d_{h+i} is trade-offed by its BERT score with the initial query q and the feedback information r_{h+i}^f:

$$i_t^f = r_{h+i} \cdot c + BERT(q, d_{h+i}) \cdot (1 - c), \tag{4}$$

where c is a coefficient (set as 0.1) to balance the influence of feedback documents and the initial query, which is widely applied in existing RF methods [7, 63]. Finally, the upcoming documents $\mathcal{D}_u$ are re-ranked with $R^{it} = \{r_{h+1}^{it}, \ldots, r_n^{it}\}$ to facilitate the search process.

4.3 RF training and evaluation

Algorithm 1 presents the training and evaluating pipelines of the RF experiments. Unlike existing literature [26, 79] that splits the dataset without considering the natural sequence of data samples, we adopt a split-by-timepoint protocol to train and evaluate the RF model. When a new search task is presented to a participant u, we first prepare the brain decoding model (shown in Algorithm 1 lines 3–4) using the brain data collected from previous search tasks of the same participants (the personalized model G_p) or other participants (the global model G_g). Training a personalized model G_p for each participant is necessary as brain signals vary across different individuals, as indicated by previous studies [39, 86]. However, a participant's data may be insufficient for training G_p at the beginning of his or her search process. Hence, we adopt a generalized model G_g trained using other participant's data as a substitute until his or her collected data size reaches a minimum required size (set as 100 data samples).

After that, experiments involving IRF and RRF are executed, as delineated in lines 5–13 and 14–18 of Algorithm 1, respectively. For IRF, valuable RF signals are gathered from historical document $\mathcal{D}_h$, enabling the generation of relevance scores for the unseen documents $\mathcal{D}_u$ utilizing the RF

ALGORITHM 1: Overall RF Pipeline and Experimental Setup

Input: A user u, a series of search tasks consist of a query set Q and document set $\mathcal{D}$ for each query $q \in Q$, a generalized brain decoding model G_g trained from other participants' brain recordings, a base ranker BERT.

Data: A personalized brain encoding model G_p share the same structure with G_g but initialized with random parameters, hyper-parameters $\Theta^{it} = \{\theta_{bs}^{it}, \theta_c^{it}, \theta_p^{it}\}$ and $\Theta^{re} = \{\theta_{bs}^{re}, \theta_c^{re}, \theta_p^{re}\}$, all collected brain data samples $\mathcal{X}_{\text{global}} = []$, all collected user annotations $\mathcal{R}_{\text{global}} = []$.

Output: The averaged document re-ranking performance S_{IRF} and S_{RRF} for IRF and RRF, respectively.

1   Initialize the brain decoding model G as G_g, and initialize both S_{IRF} and S_{RRF} as empty lists $[]$.

2   **for** *each* $q \in Q$ **do**

3   Train G_p with $\mathcal{X}_{global}$ and $\mathcal{R}_{global}$.

4   Transform G to G_p if the size of collected brain data samples $\mathcal{X}$ attains a total of 100 .

     `// Evaluating IRF performance by re-ranking unseen documents as the search proceeds.`

5   **for** *each* $h \in \{1, \ldots, h_{max}\}$`// `$h_{max}$` is the number of documents user `$u$` has interacted with in total.`

6   **do**

7   Collect user u's brain responses $\mathcal{X}$ corresponding to document $d \in \{d_1, \ldots, d_h\}$.

8   Generate R^{bs} with $\mathcal{X}$ and the brain decoding model G.

9   Generate R^c according to user u's click behaviors, calculate R^p base on the text-based ranking scores $BERT(q, D)$.

10         $R^{it} = QE^{F_{\Theta^{it}}(R^{bs}, R^c, R^p)}$. `// Generate combination relevance score for IRF.`

11   Utilize third-party relevance annotations for $d \in \{d_{h+1}, \ldots, d_n\}$ as R^{gu}.

12   S_{IRF}.append($\Pi(R^{gu}, R^{it})$). `// Calculate ranking-based metrics for IRF.`

13   **end**

     `// Evaluating RRF performance by re-ranking historical documents after the search ends.`

14   Collect user u's brain responses $\mathcal{X}$ corresponding to document $d \in \{d_1, \ldots, d_{h_{max}}\}$.

15   Generate R^{bs}, R^c, and R^p for document $d \in \{d_1, \ldots, d_{h_{max}}\}$.

16      $R^{re} = F_{\Theta^{re}}(R^{bs}, R^c, R^p)$. `// Generate combination relevance score for RRF.`

17   Collect third-party relevance annotations for $d \in \{d_1, \ldots, d_{h_{max}}\}$ as R^{gh}.

18   S_{RRF}.append($\Pi(R^{gh}, R^{re})$). `// Calculate ranking-based metrics for RRF.`

     `// Extend `$\mathcal{X}_{global}$` and `$\mathcal{R}_{global}$` for the split-by-timepoint training of brain decoding model.`

19      $\mathcal{X}_{global}$.append($\mathcal{X}$), $\mathcal{R}_{global}$.append(R^{gh}).

20   **end**

21   $S_{IRF} = Average(S_{IRF})$, $S_{RRF} = Average(S_{RRF})$

22   **return** S_{IRF}, S_{RRF};

---

method (detailed in Section 4). Given the absence of user annotations for all unseen documents, IRF performance is evaluated using third-party annotations R^{gu}, incrementally as the number of historical document h increases. Consequently, h_{max} ranking-based metrics, $\Pi(R^{gu}, R^{it})$, are calculated for a query q. In contrast, RRF performance evaluation is deferred until the end of the current search query due to its retrospective nature. The re-ranking performance for $\mathcal{D}_{h_{max}}$ is calculated by $\Pi(R^{gh}, R^{re})$, where R^{re} is the relevance score generated with the RF framework and R^{gh} indicates the user's annotation.

5   EXPERIMENTS AND RESULTS

We conduct empirical experiments (all implementation codes and datasets are available[3]) to address the following research questions:

     — **RQ1:** Can brain signals provide guidance for RF?

— **RQ2:** To what extent can we improve RF with the proposed framework and brain signals?
— **RQ3:** How do brain signals improve RF performance in different search scenarios?
— **RQ4:** Can we further improve RF by adaptively adjusting the combination weight of brain signals and other RF signals?

To address **RQ1**, we explore the effectiveness of the brain decoding model, which aims to predict brain-based relevance scores R^{bs}. Then, we analyze the document re-ranking performance of the proposed RF framework in IRF and RRF, respectively, to answer **RQ2**. Furthermore, we analyze the performance gain brought by brain signals in different search scenarios, especially in two particular cases, that is, *bad click identification* and *non-click relevance estimation* to answer **RQ3**. We observe that the benefit brought by brain signals varies with different search scenarios. Therefore, the potential advancement of RF relies on designing a better combination strategy of brain signals and other RF signals. Hence, we then explore an adaptive method to combine RF signals depending on the search scenarios, which answers **RQ4**.

5.1 Brain Decoding Experiment (RQ1)

This section elaborates on the experiments that decode brain responses to the snippet and landing page content into brain-based relevance scores. We first explain our experimental setup process, including feature extraction, decoding model selection, and evaluation protocols. Then, we show the difference between brain responses to relevant and irrelevant items with statistical analyses on the extracted features. Afterward, a decoding experiment is performed to evaluate the performance of relevance estimation based on brain signals.

5.1.1 Experimental Setup. **Feature Extraction:** We extract **differential entropy** (**DE**) [31] features for the brain decoding task. DE is a popular frequency-based feature in EEG-based prediction [24], which has shown superior performance in emotion recognition [24] and satisfaction detection [78]. The DE features are calculated in five bands, that is, δ (0.5−4 Hz), θ (4−8 Hz), α (8−13 Hz), β (13−30 Hz), and γ (30−50 Hz), on 62 EEG channels. As a result, a sample of EEG data is preprocessed into a vector of size 62×5 for each brain response upon the stimulus.

Decoding Model Selection: We adopt a **support vector machine** (**SVM**) with the Gaussian kernel as the base brain decoding model since it is prevalent and effective for brain decoding [6, 24]. Besides, it requires less computing effort than neural networks [78, 86], and can meet the demand of online training and inferencing in realistic systems. Specifically, we observe that the computing time for brain decoding with an SVM is negligible compared to (less than one-thousandth of) the BERT-based re-ranking algorithm. Hence, the proposed RF process could be conducted in realtime.

Evaluation: We evaluate the performance of the brain decoding model G (G_p or G_g) in terms of its classification abilities for relevant and irrelevant items. Since relevance annotation 1 makes up 68.2% data samples, relevance annotation 1 is regarded as a negative sample (irrelevant) and annotation 2−4 as positive samples (relevant). Hence, the classification problem is transformed into a binary classification problem, and we measure its performance with **Area Under Curve** (**AUC**). Note that we use the same classification model G (G_p or G_g) for the classification of brain response to snippet and landing page since we observe that its performance is better than training independent models.

5.1.2 Feature Analyses. Frequency-based EEG analyses are widely applied in relevance judgment [79], emotion recognition [37], and other domains. We calculate the mean value for each channel obtained by averaging DE features extracted from each participant's brain responses to web pages (both snippets and landing pages) and over two conditions: relevant and irrelevant stimuli. Figure 6 presents the F-value for 62 EEG channels by performing an ANOVA test, where the
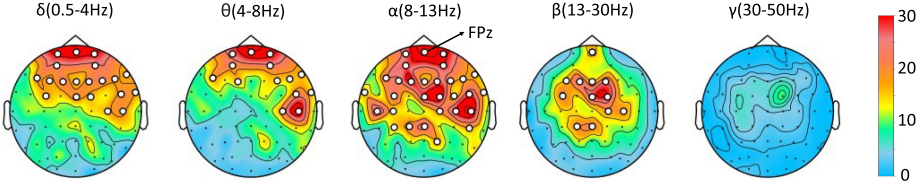
Fig. 6. Topography which shows the significance of difference (F-value) between brain response to relevant/irrelevant Web pages. Highlighted channels indicate the differences are significant at $p < 1 \times 10^{-3}$ level.

highlighted channels indicate the difference is significant at $p < 1 \times 10^{-3}$ level. We observe various significant channels in δ, θ, α, and β bands. Among these channels, the most significant difference appears in the FPz in α band ($F[1, 20] = 39.03$, $p = 4.2 \times 10^{-6}$, $M_{diff} = -0.40\ln(Hz)$). This finding indicates that brain responses to relevant and irrelevant stimuli are distinguishable.

Additionally, we observe that in δ, θ, and α bands, the most significant differences appear in the frontal region, which is inconsistent with prior literature [79]. The frontal region is related to executive functions such as judgment and problem-solving [76]. Ye et al. [79] suggests that this difference is a consequence of cognitive function and working memory executing during relevance judgment. On the other hand, we also observe significant neural differences in the central region, especially in the β and α bands. Allegretti et al. [3], Pinkosova et al. [61], Yang et al. [74] also have similar observations in the central region. Despite the different settings among these studies (stimuli based on visual [3], text [61]) and ours (stimuli with multimedia content), the common finding in the central region indicates a potential link between brain functions in this region and the concept of relevance. A partial explanation of this potential link is the memory processing during relevance judgment [61], for example, recognizing whether an item is relevant by recalling knowledge from memory.

Besides these common observations, there are also several findings in contrast with prior research. For instance, Ye et al. [79] find the most significant differences in the β band while we observe more differences in the α band than in the β band. A likely reason for this difference is the design of the search tasks: Ye et al. [79] use factoid questions, hence the participant's alert levels (reflected in β band Pfurtscheller and Da Silva [60]) will have big changes if they find the direct answer (one or two words). Unlike their study, the questions in our study are non-factoid and need the participants to judge the document's relevance with more specific content. Therefore, another major neurological phenomenon, that is, valence, may play a major role in our settings, which is also revealed in existing research [52].

5.1.3 Brain Decoding Performance. The averaged AUC performance for brain responses to landing page contents (X_l) is 0.701 ($SD = 0.059$), slightly better than that to snippet content (X_s), that is, 0.690 ($SD = 0.060$). Besides, we observe that the overall performance of the personalized model G_p (AUC=0.691, 0.681 for X_l, X_s) significantly outperforms the generalized model G_g (AUC = 0.670, 0.603 for X_l, X_s). This verifies the assumption that G_p outperforms G_g and training a personalized model for each participant is helpful for the classification performance. However, when the collected personalized data size is insufficient (for the first 100 samples), the performance of G_p (AUC = 0.584) performs worse than the general model G_g (AUC = 0.627). Hence, in the split-by-timepoint splitting protocol, it is reasonable to transform G from the generalized model G_g to the personalized model G_p. Unlike randomized data splitting, our split-by-timepoint splitting protocol closely resembles real-world scenarios to simulate the cold start situation that new users experience when they begin using our system. The results of our experiment illustrate that using a

(a) Length of time.   (b) Down-sampling rate.   (c) #Personalized data samples.
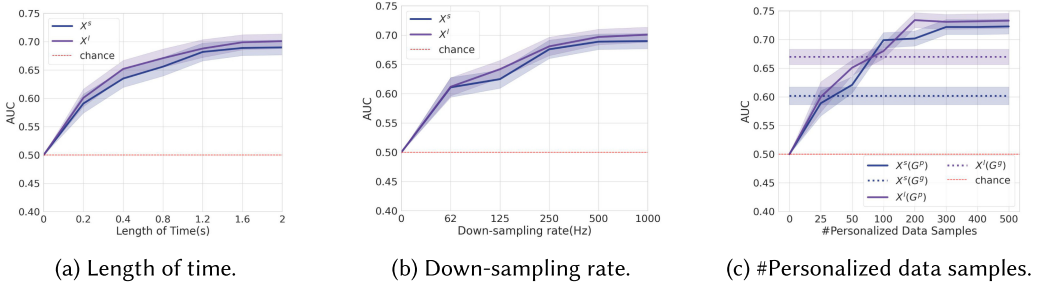
Fig. 7. AUC performance for relevance annotations of snippets and landing pages across different lengths of time, down-sampling rate, and number of personalized data samples for training. The shaded regions indicate the standard error.

general model during the cold start phase and training a subject-specific model during the search process can yield practicable results. But if some users are unwilling to provide their data for model training, utilizing general models and employing techniques like edge computing to prevent user data from being exposed can still yield satisfactory results.

As EEG data commonly contain noises, the brain decoding performance is not perfect [28, 79]. However, unlike explicit annotations, the collection of EEG signals is real-time and does not interfere with the user's search process, which makes it preferable in practice. Furthermore, there is also room for improvement in designing a sophisticated strategy to deal with the data variation problem across different individuals and devising more effective classification models. Since this is not the focus of this article, we leave the study of constructing EEG classification models in interactive search scenarios as future work.

5.1.4 Sensitivity Analysis. In Section 3.5, we elaborate on the data preprocessing protocols adopted in our study, which follows a common setting with existing neuroscientific studies [61, 79]. To explore and understand how these settings will affect the relevance prediction performance, we conduct a sensitivity analysis regarding the data-preprocessing protocols, especially the length of time segmentation and down-sampling rate, as shown in Figure 7(a) and Figure 7(b), respectively. From Figure 7(a), we can observe that the relevance estimation performance improves with the time duration increases and stabilizes between 1,600 ms to 2,000 ms. Existing literature ([3] and [79]) reported that differences in users' brain signals begin to emerge at 800 ms for relevant and irrelevant documents. However, we find there is still room for improvement with data collected over a longer length of time. On the other hand, we find that the relevance estimation performance improves if we do not down-sample too much, as shown in Figure 7(b). To reduce computational complexity, especially in real-time scenarios pertinent to RF, we adopted a consistent down-sampling rate of 500 Hz. This ensures comparable performance to scenarios without down-sampling, yet with fewer computational demands.

Additionally, we also conducted a sensitivity analysis regarding the number of training personalized data samples. In this analysis, we train the brain decoding model G^p with different sizes of collected brain data samples $\mathcal{X}$, and evaluate it with the additional 100 data samples. Furthermore, we undertake a sensitivity analysis pertaining to the number of training personalized data samples. For each participant, we train the brain decoding model G^p, using varying sizes of collected brain data samples $mathcalX$, and subsequently assess its performance using an additional 100 data samples. Figure 7(c) depicts the classification performance in terms of AUC. As shown in Figure 7(c), we observe that our model trained with cross-subject data (G^g) also exhibits good performance, which means that, we may be able to avoid using user's personal data by training and deploying

Table 3. The Statistical Results of the Mixed Linear Mode

Effects	Coef.	Std	z	p>z
Word number	0.000	0.000	1.247	0.106
Document rank	−0.007	0.004	−2.320	0.010
Task rank	−0.001	0.000	−1.542	0.062
Image size	0.001	0.000	0.099	0.921
Document relevance	0.041	0.005	8.921	0.000

Coef. and z indicate the coefficient variable and the statistic corresponding to the effect measured by brain activities, respectively.

a global brain signal model for everyone while keeping a reasonable retrieval performance. Additionally, G^p trained with a small amount of training data still performs well in the experiments (i.e., 100 data samples), which means that, in practice, we just need to collect a limited amount of user data to train the model, and then deploy the model to user's device to avoid further collection of user's personal data. These observations illustrate that we could reduce the risk of privacy-related issues in practice by using cross-subject modeling and limited sample modeling when applying BCI devices in IR scenarios.

5.1.5 Mixed Effects Analyses. In the data collection procedures, we randomized the task order and the order of the search documents, minimizing the risk of potential confounders. Despite these precautions, it remains implausible to entirely eliminate the influence of all confounding variables. Here we deliberate upon various confounding factors that could potentially influence the statistical robustness and validity of our analytical observations. Following existing literature [79], the confounding factors we considered include: individual difference (I), the task rank (O^t), the documents rank (O^d), and the word number in the documents (W), and the image size of the displayed document (S). A linear mixed model is used for modeling the dependence of brain activities (X) measured by EEG spectral powers and the ground truth relevance of documents (R^{gh}), which can be specified as follows:

$$X = (\beta_u + i_u)R^{gh} + \beta_w W + \beta_t O^t + \beta_d O^d + \beta_s S + I + \beta_0 + e, \tag{5}$$

where e represents the global residual error, β_0 denotes the global intercept, β_u, β_w, β_t, β_d, β_s denote coefficients corresponding to the effects of the confounding factors we list above. I is the individual difference effect and i_u is the coefficient for participant u. The brain activity X, estimated via spectral power as elaborated in Section 5.1, and the document relevance, R^{gh}, which spans a range from 1 to 4.

Table 3 illustrates the coefficient variables and significance of the fixed effect (i.e., the document relevance), and the random effects (i.e., word number, document rank, task rank, and image size). From Table 3, we observed that the document relevance exerted a significant effect on the brain response. This effect stood out prominently when compared to the confounding variables. Among the confounding factors, only the document rank demonstrated a significant influence on the outcome, which is reasonably attributable to the position bias effect [5, 33], where users typically perceive documents at the top of the list as more relevant. Conversely, the remaining confounding factors showed minimal effects that were not statistically significant. This suggests that while there are multiple factors influencing the brain response, the relevance of the document plays a predominant role.

*Answer to **RQ1**.* We observe that the relevance of web pages can be inferred by decoding brain signals. This demonstrates the potential of using brain signals as an additional relevance indication to supplement existing relevance signals.

5.2 IRF and RRF Experiment (RQ2)

5.2.1 Experimental Setup. **Evaluating Protocols:** In IRF, we evaluate the re-ranking performance of the upcoming documents $\mathcal{D}_u = \{d_{h+1}, d_{h+2}, \ldots, d_n\}$ in an interactive manner: calculate and average the metric $\Pi(R^{gu}, R^{it})$ as h increases from 1 to h_{max}. As we do not require the participants to annotate the relevance for unseen documents $\mathcal{D}_u$, we use the external annotations as the ground truth relevance R^{gu}. In RRF, we calculate the metric $\Pi(R^{ge}, R^{re})$ of the historical documents $\mathcal{D}_h = \{d_1, d_2, \ldots, d_h\}$ for $h = h_{max}$ and adopt the user's annotation to obtain R^{ge}. Since the landing page contains more content about the document than the snippet content, the document's ground truth relevance is assigned as the landing page's annotation if the document is clicked. Otherwise, if the document is non-clicked, we simply use the annotation of the snippet as a substitution. For ranking-based evaluation metric Π, we adopt **Mean Average Precision (MAP)** and NDCG at different cutoffs: 1, 3, 5, and 10 [16]. A two-tailed t-test is then applied to measure the significance of the re-ranking performance achieved by different methods and signals.

Baselines: For IRF, the baselines include three re-ranking strategies without any user signals: BM25 [62], BERT(R^p) [58] (re-rank according to the BERT re-ranker, equivalent to re-ranking by R^p), and Sogou (using the original ranking in the Sogou search engines). Besides, we report the performance of our proposed BERT-based query expansion method $QE^{F_{\theta^{it}}(R^{bs}, R^c, R^p)}$ and its ablations (QE^{R^p}, $QE^{F_{\theta^{it}}(R^c, R^p)}$, and $QE^{F_{\theta^{it}}(R^{bs}, R^p)}$). In addition to the proposed RF method, we also report the performance of a traditional RF method RM3 [40]. The implementation of RM3 inherits parameters from Lavrenko and Croft [40] and selects the number of rewriting terms from $\{3, 5, 10\}$. For RRF, the baselines also include three re-ranking strategies without user signals: BM25, BERT, and Sogou. In addition to the proposed framework, which combines relevance scores from all signals ($F_{\theta^{re}}(R^{bs}, R^c, R^p)$), we also report its ablations $F_{\theta^{re}}(R^c, R^p)$ (without brain signals) and $F_{\theta^{re}}(R^{bs}, R^p)$ (without click signals).

Selection of Combination Parameters: Each combination parameter $\theta^{*,\dagger}$ ($* \in \{it, re\}, \dagger \in \{bs, c, p\}$) is selected from $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. In our experiments, we first explore the overall document re-ranking performance of our framework with fixed values of Θ^{it} and Θ^{re}. To initialize the fixed parameters, we randomly sample a subset consisting of 200 search tasks and test the document re-ranking performance in terms of NDCG@10. Then the optimal parameters among all the combinations (Θ^{it} in IRF, Θ^{re} in RRF) on this subset are selected, which are $\theta^{it, bs:c:p} = 3 : 1 : 1$ in IRF, and $\theta^{re, bs:c:p} = 5 : 2 : 0$ in RRF. We observe that the chosen parameters also achieve the best performance in the whole dataset in comparison with other parameters, which indicates that the chosen parameters are robust. Note that similar to most existing RF frameworks [63, 64, 81], our experiment does not involve a large number of training parameters. Instead, we only tune the combination parameters to merge different RF signals and use a pre-trained retrieval model for document re-ranking.

As $\theta^{re, p}$ is 0 in the selection of parameters for RRF, we further search in the 0–0.2 range and evaluate the performance in the whole dataset. We observe that the performance of $F_{\Theta^{re}}(R^{bs}, R^c, R^p)$ yields no significant difference in the range of 0–0.14 (peaks at 0.06). However, the performance of $BERT(R^p)$ is significantly higher than chance (i.e., rank the documents randomly) with $p < 0.001$. This indicates that pseudo relevance is effective for RRF, but when compared to the user signals (R^c and R^{bs}), the effect is very small. As the documents in our experiment are selected from top documents while submitting the query to Sogou's search engine, as explained in Section 3.2. Therefore, most of them are already semantically related to the query term and their semantic relevance to the query is similar, which is pseudo-relevance measures. Therefore, knowing what topics the user needs is more important than the little differences in semantic relevance, given that the queries

Table 4. The Document Re-Ranking Performance in IRF

Method[1]	NDCG@1	NDCG@3	NDCG@5	NDCG@10	MAP
BM25	0.2031^*	0.2265^*	0.2449^*	0.3031^*	0.3048^*
Sogou	0.2085^*	0.2284^*	0.2481^*	0.3065^*	0.3214^*
BERT (R^p)	0.2060^*	0.2326^*	0.2579^*	0.3221^*	0.3033^*
$QE^{F_{\theta^{it}}(R^p)}$	0.2306^*	0.2361^*	0.2582^*	0.3205^*	0.3279^*
$QE^{F_{\theta^{it}}(R^{bs},R^p)}$	0.2477^*	0.2569^*	0.2774^*	0.3374^*	0.3235^*
$QE^{F_{\theta^{it}}(R^c,R^p)}$	0.2842^*	0.2952^*	0.3124^*	0.3690^*	0.3708^*
$RM3^{F_{\theta^{it}}(R^{bs},R^c,R^p)}$	0.2332^*	0.2523^*	0.2717^*	0.3289^*	0.3337^*
$QE^{F_{\theta^{it}}(R^{bs},R^c,R^p)}$	**0.2948**	**0.3024**	**0.3191**	**0.3747**	**0.3744**

*indicates a significant difference in performance when compared to $QE^{F_{\theta^{it}}(R^{bs},R^c,R^p)}$, significant at a level of $p < 1 \times 10^{-3}$.
[1]$R^\dagger$ indicates relevance score based on signals †. bs, c, and p indicate brain signals, click signals, and pseudo-relevance signals, respectively.
Results in boldface denote the winning performance for each column.

in our tasks are short and with broad topics. We acknowledge the limitation of this selection of queries and further discuss it in Section 6.3.

In addition, we also explore the performance of different combination parameters in Section 5.3 to show how we should re-weight the importance of various RF signals regarding the search scenarios. Motivated by the findings in Section 5.3, we further explore an adaptive RF signals combination method in Section 5.4 which combines the importance of RF signals depending on the search scenario. The adaptive signals combination method achieves better performance improvement than using fixed parameters in the IRF task.

5.2.2 Overall Results. **IRF Performance.** Table 4 presents the document re-ranking performance in IRF. From Table 4, we have the following observations: (1) Methods using feedback information gathered from user interactions (i.e., $QE^{F_{\theta^{it}}(R^{bs},R^p)}$, $QE^{F_{\theta^{it}}(R^c,R^p)}$, and $QE^{F_{\theta^{it}}(R^{bs},R^c,R^p)}$) outperform methods without regard to user signals (i.e., BM25,BERT(R^p), Sogou, and QE^{R^p}). Although existing research underscores the superior performance of BERT over BM25 [58], the divergence in their performance becomes notably minimal when juxtaposed with the impact of integrating additional feedback information, especially brain signals. This indicates that personal factors extracted from user interactions are helpful for improving document re-ranking performance. (2) With additional relevance score R^{bs} extracted from brain signals, the query expansion (QE) method receives a significant performance boost. The difference between the performance (in terms of NDCG@10) of $QE^{F_{\theta^{it}}(R^{bs},R^p)}$ ($QE^{F_{\theta^{it}}(R^{bs},R^c,R^p)}$) and QE^{R^p} ($QE^{F_{\theta^{it}}(R^c,R^p)}$) is 5.3% (1.5%), which is significant at $p = 7.8 \times 10^{-16}$ (1.5×10^{-6}) with a pairwise t-test. This demonstrates that brain signals can provide additional information to existing signals (pseudo-relevance signals or a combination of pseudo-relevance signals and click signals). (3) The proposed BERT-based query expansion method is more effective than the conventional method RM3. This may suggest that BERT is better at capturing a document's semantic representation than statistical LM.

RRF Performance. Table 5 presents the document re-ranking performance in RRF. We observe similar findings as we have discussed in IRF. First, models that use user signals (i.e., $F_{\theta^{re}}(R^{bs},R^p)$, $F_{\theta^{re}}(R^c,R^p)$, and $F_{\theta^{re}}(R^{bs},R^c,R^p)$) often have better performance than those that do not (i.e., BM25, BERT and Sogou). Second, brain signals can boost RF performance, for example, $F_{\theta^{re}}(R^{bs},R^c,R^p)$ leads to a performance gain of 7.4% in terms of NDCG@10 over $F_{\theta^{re}}(R^c,R^p)$.

Table 5. The Document Re-Ranking Performance in RRF

Method[1]	NDCG@1	NDCG@3	NDCG@5	NDCG@10	MAP
BM25	0.3113*	0.3831*	0.4703*	0.5587*	0.5477*
BERT (R^p)	0.3218*	0.3846*	0.4626*	0.5605*	0.5534*
Sogou	0.3313*	0.4046*	0.4940*	0.5848*	0.5633*
$F_{\Theta^{re}}(R^{bs}, R^p)$	0.5074*	0.5494*	0.6254*	0.6973*	0.6744*
$F_{\Theta^{re}}(R^c, R^p)$	0.5426*	0.5936*	0.6578*	0.7161*	0.6694*
$F_{\Theta^{re}}(R^{bs}, R^c, R^p)$	**0.6350**	**0.6617**	**0.7171**	**0.7693**	**0.8009**

*indicates the difference in performance, when compared to $F_{\Theta^{re}}(R^{bs}, R^c, R^p)$, is significant at a level of $p < 1 \times 10^{-3}$.

[1]$R^{\dagger}$ indicates relevance score based on signals †. bs, c, and p indicate brain signals, click signals, and pseudo-relevance signals, respectively.

Results in boldface denote the winning performance for each column.

These similar findings indicate that in both IRF and RRF, brain signals can be utilized to improve document re-ranking performance.

Except for the similarities, we also notice differences between IRF and RRF. For example, we observe that the performance gain achieved by adding user signals (i.e., brain signals or click signals or their combination) is larger in RRF than in IRF. Especially, in RRF, the performance gain between $F_{\theta^{re}}(R^{bs}, R^c, R^p)$ and $F_{\theta}^{re}(R^c, R^p)$ ($F_{\theta^{re}}(R^{bs}, R^c, R^p)$ and R^p) is 7.4% (37.3%) in terms of NDCG@10. However, in IRF, the performance gain between $QE^{F_{\theta^{it}}(R^{bs}, R^c, R^p)}$ and $QE^{F_{\theta^{it}}(R^c, R^p)}$ ($QE^{F_{\theta^{it}}(R^{bs}, R^c, R^p)}$ and QE^{R^p}) is smaller, that is, 1.5% (16.9%) in terms of NDCG@10. One possible reason is the fact that ranking unseen documents is much more difficult than ranking historical documents. With accurate RF signals, we can easily create the best ranking of historical documents, but may still struggle in estimating the relevance of new documents because there is no guarantee that the new relevant documents would be similar to the historical relevant documents we observed. However, in Section 5.3, we compare the potential improvements brought by ideally combining different RF signals and observe a larger potential in IRF. Then in Section 5.4, we propose a combination method that helps IRF exhibit comparable performance to RRF.

*Answer to **RQ2***. We verify that the RF performance can be significantly improved when adding brain signals into the proposed framework, that is, a performance gain of 7.4% and 1.5% in terms of NDCG@10 in IRF and RRF, respectively.

5.3 In-Depth Analyses (RQ3)

This section explores the RF framework based on **combination parameter analyses** and **search scenario analyses**. The combination parameter analyses utilize an ideal experiment to examine the potential of improving RF performance by selecting ideal combination parameters in different scenarios. It is observed that potential advancement lies in designing better combination weights for RF signals, especially in IRF. On the other hand, the search scenario analyses present the gain of brain signals on RF performance in several specific search scenarios. The analyses further show that it is necessary to increase the combination weight of brain signals in those search scenarios. The analyses in these two aspects reveal that RF performance improvement lies in designing better combination protocols, which is elaborated in Section 5.4.

5.3.1 Combination Parameter Analyses. In the above experiment, we used a fixed selection of combination parameters Θ^{it} and Θ^{re} which has averagely the best performance (detailed in Section 5.2.1). However, there is still potential for improvement in using different parameters to
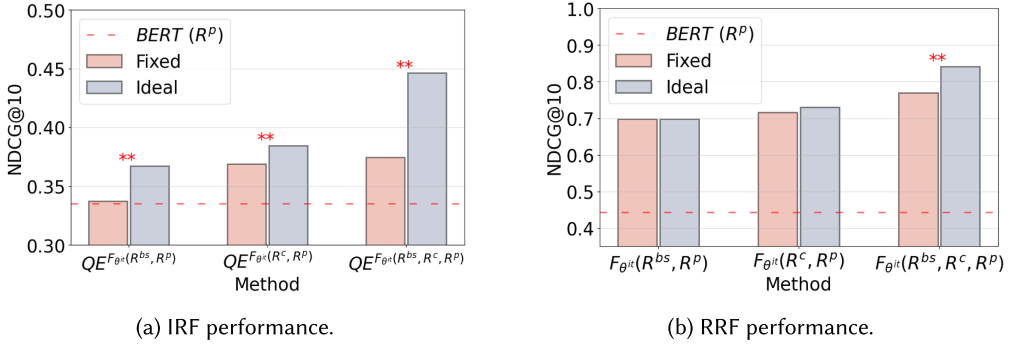
(a) IRF performance.

(b) RRF performance.

Fig. 8. The RF performance with fixed and ideal combination parameter Θ. ** indicates significant differences between RF methods using fixed and ideal combination parameters at $p < 0.01$ level using a pair-wise t-test.

combine RF signals for each data sample. Hence, we conduct an ideal experiment, which adaptively selects the ideal combination parameters Θ^{it} and Θ^{re} to achieve the best document re-ranking performance. For every data sample, corresponding to each computation in Algorithm 1 (line 10 utilizing Θ^{it} and line 16 employing Θ^{re}), we search values for $\theta^{\cdot,\dagger}$, where $\in it, re$ and $\dagger \in bs, c, p$, from the set 0.0, 0.2, 0.4, 0.6, 0.8, 1.0. Subsequently, we determine the optimal combination of parameters Θ^{it} and Θ^{re} based on the NDCG@10 for each distinct data sample. Note that the performance of the ideal experiment is not achievable in reality. This ideal experiment aims to compare the difference between the practical experiment results and the ideal results, which indicates the potential of the RF framework and brain signals.

Figure 8 presents the RF performance with fixed and ideal combination parameter Θ. From Figure 8, we observe that RF with the ideal combination parameters outperforms RF with the fixed combination parameters in all methods in both IRF and RRF tasks. This is obvious since the selection of the ideal combination parameter is directly based on the RF performance. Besides, we observe that the performance difference of RF with the fixed/ideal combination parameters is larger in IRF than in RRF. Especially, the performance improvement in IRF achieves 19.1% (0.4463 in comparison with 0.3747 in terms of NDCG@10) with the method of $QE^{F_{\theta^{it}}(R^{bs}, R^c, R^p)}$, which is significant at $p < 1 \times 10^{-2}$. This indicates that IRF exhibits greater potential than RRF when effectively combining brain signals, click signals, and pseudo-relevance signals.

A potential explanation for this difference between IRF and RRF is their different task formulation and measurements. In RRF, the combination score is directly applied to re-rank the feedback documents (denoted as $\mathcal{D}_h$ in the task formulation). Hence, the combination score only needs to reflect the differences between each other and can be used for ranking purposes. However, in IRF, the combination score is applied to a query expansion module. Therefore, it is not only necessary to maintain the ranking of documents' relevance, but also to have reasonable values determine their weight in query rewriting. For example, we assume that d_1 and d_2 are two feedback documents and d_1 is more relevant to the user's intent. In RRF, it is no problem to assign any relevance score to d_1 and d_2 which only needs to maintain that the score of d_1 is higher than that of d_2. However, in IRF, relevance scores of d_1 and d_2 directly reflect the importance of d_1 and d_2 in query expansion. Hence, setting proper relevance scores for them is much more important to avoid possible bad cases, for example, d_2 may even be harmful to query expansion.

5.3.2 Search Scenario Analyses. The RF performance in search scenarios where click signals are missing or biased is explored in this section. It is observed that brain signals are particularly
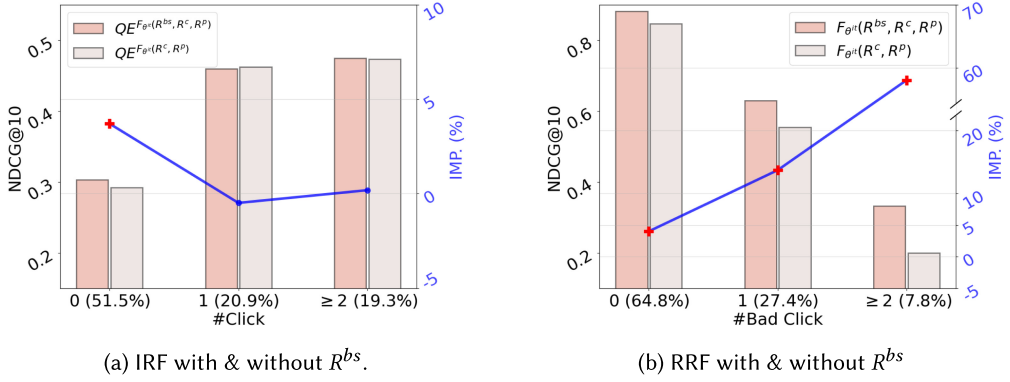
Fig. 9. The RF performance with & without brain signals in different search scenarios. + denotes the improvement (IMP.) is significant ($p < 0.05$) using a pairwise t-test.

beneficial when clicks are missing (analyzed in IRF as kicking off RF before any click happens is significant in an interactive process) by estimating non-click relevance. Besides, brain signals help cases where bad clicks happened (analyzed in RRF so that we can re-rank bad clicks for potential search processes in the future).

Non-click Scenarios. There are good reasons not to treat all non-clicks equally, for example, some non-clicks may contain useful information [79] and even attract revisiting behaviors [73]. In our user study, we also observe that the non-click documents are annotated with different relevance scores (1 (77.6%), 2 (7.4%), 3 (4.6%), and 4 (10.2%)). Based on the above observation, we aim to explore whether brain signals can estimate non-click relevance and kick off IRF before any click happens. In the brain decoding experiment, we observe that the AUC performance for non-click data samples achieves 0.675, which presents a slight decrease compared to that measured in all data samples but is still significantly better than random. This motivates us to further explore whether brain signals can improve RF performance when click signals are missing.

Figure 9(a) presents the IRF performance in search scenarios with different numbers of clicks. From Figure 9(a), we observe that the performance gain between $QE^{F_{\theta^{it}}(R^{bs}, R^c, R^p)}$ and its ablation $QE^{F_{\theta^{it}}(R^c, R^p)}$ is largest in a non-click search scenario (an improvement of 3.7% in terms of NDCG@10). In contrast, their performance differences are not significant in search scenarios with not less than one click. As IRF is an interactive process, the number of clicks increases as the number of examined documents increases (Pearson's $r = 0.45$ ($p < 1 \times 10^{-3}$)). Hence, we also observe that the performance difference between $QE^{F_{\theta^{it}}(R^{bs}, R^c, R^p)}$ and $QE^{F_{\theta^{it}}(R^c, R^p)}$ is larger at the beginning of the search process (an improvement of 3.3% in terms of NDCG@10 for session length $h \leq 4$) than the subsequent search process (an improvement of 0.6% in terms of NDCG@10 for session length $h > 4$). As the non-click search scenario accounts for a large proportion (51.5%) and it is usually more important to quickly improve retrieval performance at the beginning of the search process, we argue that brain signals have great potential in IRF.

Furthermore, we analyze how the IRF performances vary with the selections of combination parameter Θ^{it}. We select $\theta^{it,bs:c:p}$ from {3:1:1, 5:1:1, 1:5:1}, where 3:1:1 is the averagely optimal parameter among all data samples while 5:1:1 and 1:1:5 are combinations that emphasize the importance of brain signals and click signals, respectively. As shown in Figure 10(a), in a non-click search scenario, combining brain signals with a higher weight ($\Theta^{it,bs:c:p} = 5:1:1$) outperforms other combination parameters $\Theta^{it,bs:c:p} = 3:1:1$ (1:5:1) significantly at $p = 2.6 \times 10^{-10}$ (6.3×10^{-17}) using a pair-wise t-test. Besides, if we simply set the parameter Θ^{it} as 5:1:1 in a non-click scenario, and 3:1:1

(a) IRF with different Θ^{it}.
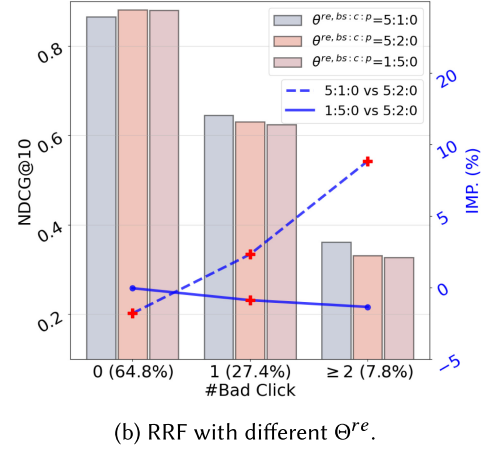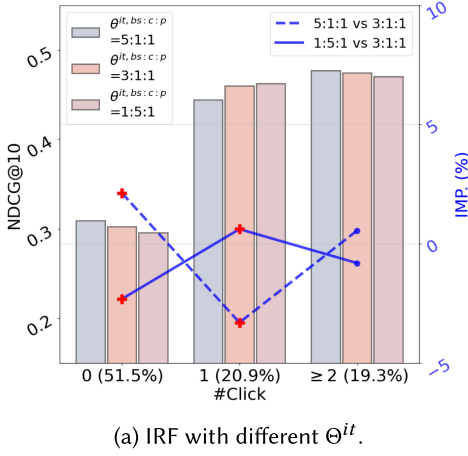
(b) RRF with different Θ^{re}.

Fig. 10. The RF performance with different combination parameter Θ in various search scenarios. + denotes the improvement (IMP.) is significant (p < 0.05).

in others, the IRF performance in terms of NDCG@10 can achieve 0.3781, significantly performing better than the performance achieved by using fixed combination parameters ($p = 2.7 \times 10^{-10}$). This reveals the potential benefit if we adaptively integrate brain signals regarding the search context.

Bad Click scenarios. Bad click indicates a document clicked by the user is irrelevant and may lead to a poor search experience. This usually happens if the document's snippet is attractive, but its landing page content is unsatisfactory [46]. In our experiment, we define a "bad click" if the relevance annotation of the landing page is 1 ("totally irrelevant") or 2 ("irrelevant") for a clicked document, and we observe that a proportion of 21.8% clicks are grouped into "bad click". Since brain signals are effective in inferring the relevance of the landing pages (with a binary classification AUC = 0.703), it is interesting to further explore whether brain signals can boost RRF performance in cases where a bad click happens.

As presented in Figure 9(b), we observe that the performance difference between $F_{\theta^{re}}(R^{bs}, R^c, R^p)$ and $F_{\theta^{re}}(R^c, R^p)$ is larger as the number of bad clicks increase. This indicates brain signals can bring more benefits in scenarios where bad clicks often happen. Furthermore, we explore combination parameters Θ^{re} that averagely performs best ($\theta^{re,bs:c:p} = 5:2:0$) and emphasize the importance of brain signals ($\theta^{re,bs:c:p} = 5:1:0$) and click signals ($\theta^{re,bs:c:p} = 1:5:0$), respectively. From Figure 10(b), we observe that combination with $\theta^{re,bs:c:p} = 5:1:0$ performs much better than using the averagely optimal parameter $\theta^{re,bs:c:p} = 5:2:0$. This emphasizes the need to prioritize brain signals in cases where there may be negative or improper clicks. If we simply adopt parameter $\theta^{re,bs:c:p} = 5:1:0$ in search scenarios where at least one bad click happens and 5:2:0 in others, the RRF performance in terms of NDCG@10 can achieve 0.7756 (significantly better than using fixed parameters at $p = 1.0 \times 10^{-3}$ using a pair-wise t-test). Although the number of bad clicks is actually not available in practice, this observation reveals a possibility to better combine brain signals into RF. The studies of detecting search scenarios that potentially lead to bad click and designing adaptable combination strategies are left as future work.

5.3.3 Case Study. Table 6 presents an example of IRF and RRF results for participant ID 1 and query "The Prophet" with task description "explore the concepts of prophet in the general domain". In this scenario, the participant goes through historical documents ranging from d1 to d6. For the IRF task, the objective is to re-rank the unseen documents (D^u), d7 through d12,

Table 6. Example of IRF and RRF Results for Participant ID 1 and Query "The Prophet" with Task Description "Explore the Concepts of Prophet in the General Domain"

(a) The titles (translated into English) for historical documents (D^h) and unseen documents (D^u).

Query: The Prophet	
Historical documents (D^h)	Unseen documents (D^u)
d1: The Prophet - French, rough, crime	d7: Analysis and translation of the Prophet's poem
d2: Miracles of the Mohammed Prophet	d8: What does the prophet mean in the Bible? - Sogou ask
d3: The Prophet poem by Gibran	d9: The Prophet movie: free online resource
d4: The Prophet - Sogou encyclopedia	d10: The Prophet : Gibran
d5: Fifth personality - introduction of the Prophet	d11: The Prophet in different cultures
d6: General concepts of "Prophet"	d12: The Prophet - Chinese dictionary

(b) The Estimated Base Relevance Scores for Historical Document D^h, R^c, R^p, and R^{bs} Represent Click-Based, Pseudo-Relevance Based, and Brain-Based Scores, Respectively

D^h	R^c	R^p	R^{bs}
d1	0	0.6	0.3
d2	0	0.3	0.6
d3	0	0.4	0.3
d4	1	0.4	0.7
d5	0	0.3	0.2
d6	1	0.5	0.6

(c) The Re-Ranked Documents List for IRF & RRF Tasks Across RF Models with and Without Brain Signals (w R^{bs} and w/o R^{bs})

IRF	w R^{bs}	d11,d12,d8,d7,d9,d10
	w/o R^{bs}	d12,d11,d8,d7,d9,d10
RRF	w R^{bs}	d4,d6,d2,d1,d3,d5
	w/o R^{bs}	d6,d4,d1,d3,d2,d5

Documents with relevance annotations of 4 and 2-3 are highlighted in purple and light purple, respectively.
Documents with annotations of 1 are presented in black.

whereas the RRF task focuses on re-ranking the historical documents(D^h) from d1 to d6. Table 6a presents the document titles. Under the task description, document $d4$, $d6$, and $d11$ are highly relevant (with relevance annotation 4). On the other hand, documents $d2$ and $d8$ are partially relevant since they are related to the prophet in Islam and Christian, respectively. Both are subsets of the broader concept of the prophet. Documents $d12$ is also partially relevant since the dictionary explanation of "Prophet" may provide some useful information, though it is not concrete enough.

Table 6b presents the estimated base relevance scores for historical documents D^h. From Table 6b, we can observe that R^p does not align with the ground truth relevance very well, as it is only based on the query terms encompassing broad topics. On the other hand, R^c and R^{bs} are aligned with the ground truth relevance. Additionally, R^{bs} can provide additional information especially when two documents are both non-click or clicked. As shown in Table 6c, in the RRF task, d1,d2,d3, and d5 are non-clicks. But RF with R^{bs} accurately ranks d2 at the top of them because the user's brain response provides additional information of non-clicks. On the other hand, the user's brain response provides information that the user is more satisfied with d6 regarding the clicked documents d4 and d6. Hence, in the IRF task, RF with R^{bs} ranks d11 ahead of d12 because d11 is semantically more close to the most satisfied historical document (i.e., d6) than d12.

*Answer to **RQ3**.* We demonstrate that brain signals are particularly helpful in search scenarios involving non-click relevance estimation (in IRF) and bad click identification (in RRF). Therefore, it is imperative to prioritize brain signals in search scenarios where click signals may be biased.

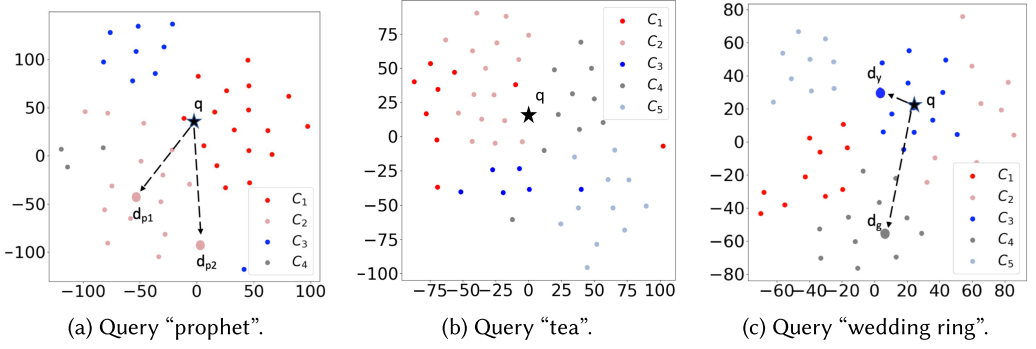(a) Query "prophet". (b) Query "tea". (c) Query "wedding ring".

Fig. 11. T-SNE plot of BERT embeddings of queries and their corresponding documents with different intents (painted in different colors). The IRF process utilizes a query expansion module to bring the query representation closer to the user's current search intent.

5.4 Adaptive Combination of RF Signals (RQ4)

In Section 5.3, we have demonstrated that the importance of different RF signals varies with the search scenarios. This indicates the potential for an adaptive RF signal combination strategy. Particularly, as shown in Figure 8, IRF has a larger improvement space than RRF since the performance difference between the fixed signal weighting and the ideal signal weighting is huge. Motivated by these observations, we propose and experiment with an adaptive RF signals combination method by synthesizing user signals to learn adaptive combination weights in different search scenarios.

5.4.1 Modeling Motivation. We illustrate and summarize the modeling motivation of the adaptive RF signals combination method. To facilitate ease of understanding, this section inherits the notations presented in Table 2.

As discussed in Section 5.3, setting proper combination parameters is more significant in IRF than in RRF. Hence, we focus on applying the adaptive RF signals combination method in IRF in this section. Before introducing the motivation of the adaptive RF signals combination method, we present an illustration of the IRF task in Figure 11. The IRF process is designed to bring the query representation closer to the user's current search intent, as queries submitted to the search engine are usually broad and may be related to different subtopics. For instance, as depicted in Figure 11(a), the documents corresponding to the query "prophet" are categorized into four clusters C_1, C_2, C_3, and C_4, each representing a different subtopic. Suppose the user's search intent is linked to the documents in cluster C_2 and shows a high RF score on document d_{p1}. In that case, the IRF will bring the original query representation closer to d_{p1}. Thus, other documents (e.g., d_{p2}) in cluster C_2 would acquire higher relevance scores than documents in other clusters.

We denote a possible search scenario as $S_c = \{q, \mathcal{D}_h, \mathcal{D}_u, n_h\}$ in which the user has examined h documents under query q, and the number of user's clicks is n_h ($n_h \in \{1, 2, \ldots, h\}$). Then, we explain why the combination parameter $\Theta^{it,S_c} = \{\theta^{it,bs,S_c}, \theta^{it,c,S_c}, \theta^{it,p,S_c}\}$ should be specially selected for different search scenarios S_c. First, to effectively utilize the documents in $\mathcal{D}_h$ as feedback signals, the documents' representativeness within their respective clusters should be considered. For example, as shown in Figure 11(a), both document d_{p1} and document d_{p2} belong to the same document cluster C_2. However, d_{p1} is more representative than d_{p2} as it is closer to the average representation of documents in C_2. Hence, the feedback information of d_{p1} is more valuable than d_{p2} as it is more representative among documents in C_2. To boost RF performance, when d_{p1} (d_{p2}) belongs to the feedback documents $\mathcal{D}_h$, Θ^{it,S_c} should be assigned with a higher (lower) weight. Second, the click probability of a document is influenced not only by its relevance but also by

---

ALGORITHM 2: Adaptive RF Signals Combination

Input: A search scenario $Sc = \{q, \mathcal{D}_h, \mathcal{D}_u, n_h\}$, where n_h is selected from $\{1, 2, \ldots, h\}$; Documents clustering $\mathcal{D} = \{C_1, C_2, \ldots, C_{qm}\}$. Synthesis times T.

Data: All candidate combination parameters $\Theta^{it} = \{\theta^{it,bs}, \theta^{it,c}, \theta^{it,p}\}$, where $\theta^{it,bs}, \theta^{it,c}, \theta^{it,p}$ are selected from $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$; Best combination parameter $\Theta^{it,Sc}$ for search scenario Sc, initialized as $\{0, 0, 0\}$; Best synthesized performance $\hat{\Pi} = 0$.

Output: The adaptive RF combination parameters $\Theta^{it,Sc}$.

1 **for** *each* Θ^{it} **do**
2   Sum of performance $\Pi = 0$;
3   **for** *each* $C_j \in \{C_1, C_2, \ldots, C_{qm}\}$ **do**
4   Synthesize T possible click-based and brain-based relevance scores $\{R^{c,Sc}, R^{bs,Sc}\}$ in Sc when assuming documents in C_j as relevant.
5   Calculate the averaged RF performance Π_{C_j} with the synthesized relevance scores $\{R^{c,Sc}, R^{bs,Sc}\}$ and candidate combination parameter Θ^{it}.
6         $\Pi = \Pi + \Pi_{C_j}$;
7   **end**
8   **if** $\Pi > \hat{\Pi}$ **then**
9         $\Theta^{it,Sc} = \Theta^{it}$; $\hat{\Pi} = \Pi$;
10   **end**
11 **end**
12 **Return** $\Theta^{it,Sc}$;

---

other independent factors, for example, some irrelevant documents may attract users' clicks due to the "clickbait" issue. Hence, whether the click signals are reliable and how should we balance the combination weights of click signals and brain signals (i.e., balance θ^{it,bs,S_c} and θ^{it,c,S_c}) should depend on the search scenario $\mathcal{S}_c$. Finally, the distance between documents in $\mathcal{D}_h$ and the original query q varies with different documents. For example, as shown in Figure 11(c), d_y is more similar to the original q than d_g. Hence, the weight of θ^{it,p,S_c} should be set differently depending on the representation difference between the documents in feedback documents $\mathcal{D}_h$ and the original query q.

5.4.2 Overall Pipeline. As shown in Algorithm 2, the adaptive RF signals combination method generates adaptive combination parameters Θ^{it,S_c} for a possible search scenario $\mathcal{S}_c = \{q, \mathcal{D}_h, \mathcal{D}_u, n_h\}$. The method includes the following process: First, we cluster the documents $\mathcal{D}$ corresponding to the query q into q_m clusters $\{C_1, C_2, \ldots, C_{q_m}\}$. Second, for each document cluster N, we assume that documents within this cluster are related to the user's search intent and synthesize the user's possible click-based and brain-based relevance scores $\{R^{c,S_c}, R^{bs,S_c}\}$ under this assumption. Finally, with the synthesized user signals $\{R^{c,S_c}, R^{bs,S_c}\}$, we search for the combination parameters Θ_{it} with the averagely best RF performance, which is denoted as Θ^{it,S_c}. With the above process, we generate the adaptive RF combination parameters Θ^{it,S_c} that can be applied to combine the actual user signals during the search process.

5.4.3 Preparation. We cluster the documents $\mathcal{D}$ corresponding to the query q into q_m clusters $\{C_1, C_2, \ldots, C_{q_m}\}$ following Liu et al. [43]'s method. Liu et al. [43] assume that a query may contain several subtopics and adopt a multi-step procedure to generate the subtopics and classify each document into one of the subtopics with manual effort. Examples of the document clustering results are presented in Figure 11, in which each document is visualized with its BERT embeddings. As the document clustering method is not the concentration of this article, we leave the exploration of

other automatic subtopic mining baselines with competitive performance or unsupervised settings as future work.

5.4.4 Click-Based and Brain-Based Relevance Score Synthesis. For a search scenario $S_c = \{q, \mathcal{D}_h, \mathcal{D}_u, n_h\}$, we synthesize the user's possible click-based and brain-based relevance scores R_{c,S_c} and R^{bs,S_c} for each document in $\mathcal{D}_h$. We assume that query q is related to q_m search intents $\{Si_1, \ldots, Si_{q_m}\}$ in which Si_j is related to one of the cluster C_j ($j \in \{1, \ldots, q_m\}$). Then we iterate over every cluster C_j and synthesize N (set as 20) user behaviors on each iteration. The synthesis will generate click-based relevance scores R^{c,S_c} and brain-based relevance scores $R^{bs,Sc}$ for each scenario Sc. Note that we simply assume that all search intents Si_j corresponding to a given query have uniform possibilities. Hence, the synthesis times are set to the same number N for each document cluster C_j. This could be changed if we have more prior information regarding the distribution of different search intents.

When iterating on the jth cluster C_j, the click-based relevance score for the ith document, denoted as $r_i^{c,Sc}$, is synthesized following the Bernoulli distribution with a constraint of their sum is n_h:

$$\forall i, r_i^c, Sc \sim \begin{cases} Bernoulli(p_{c,rel}), & \text{if} \quad d_i \in C_j \\ Bernoulli(p_{c,irel}), & \text{if} \quad d_i \notin C_j \end{cases} \quad \text{subject to} \sum_i^h r_i^{c,sc} = Ci_h, \quad (6)$$

where $p_{c,rel}$ ($p_{c,irrel}$) is a parameter inferred from the distribution of the brain-based relevance scores in the user study's relevant (irrelevant) documents with an interval estimation, that is, $p_{c,rel}$ ($p_{c,irrel}$) indicates the possibility that a relevant (irrelevant) document is clicked. The synthesis ensures that each $r_i^{c,Sc}$ follows the Bernoulli distribution, and the synthesized total number of clicks in $\mathcal{D}_h$ is n_h. This constraint is due to the number of total clicks being n_h in a given search scenario Sc. To solve this constrained distribution in practice, we simply keep synthesizing a series of $r_i^{c,Sc}$ ($i \in \{1, \ldots, h\}$) following the Bernoulli distributions until their sum is n_h.

On the other hand, the ith document d_i's brain-based relevance score r_i^{bs} is synthesized following a normal distribution:

$$r_i^{bs} \sim \begin{cases} Normal(\mu_{bs,rel}, \sigma_{bs,rel}), & \text{if} \quad d_i \in C_j \\ Normal(\mu_{bs,irel}, \sigma_{bs,irel}), & \text{if} \quad d_i \notin C_j \end{cases}, \quad (7)$$

where $\mu_{bs,rel}$ and $\sigma_{bs,rel}$ ($\mu_{bs,rel}$ and $\sigma_{bs,irel}$) are parameters inferred from the distribution of the brain-based relevance scores in the user study's relevant (irrelevant) documents with an interval estimation.

5.4.5 Optimal Combination Parameters Searching. After the synthesis process, for a search scenario $S_c = \{q, \mathcal{D}_h, \mathcal{D}_u, n_h\}$, we synthesize N corresponding R_c and R^{bs} for every document cluster C_j ($j \in \{1, \ldots, q_m\}$). Then we search for the optimal combination parameters $\Theta^{it,Sc}$ and $\Theta^{re,Sc}$, respectively. For all selections of parameters, that is, $\theta^{it,bs,Sc}$, $\theta^{it,c,Sc}$, and $\theta^{it,p,Sc}$ selected from $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, we compute the ranking-based evaluation metrics $\Pi(R^{gu}, R^{it})$, where R^{gu} is evaluated by assuming documents belong to C_j ($j \in \{1, \ldots, q_m\}$) as relevant:

$$r_i^{gu} = \begin{cases} 1, & \text{if} \quad d_i \in C_j \\ 0, & \text{if} \quad d_i \notin C_j \end{cases}. \quad (8)$$

Then we select the optimal combination parameters which achieve the averagely best RF performance (Π is set as $NDCG@10$ in our experiment) among all $N \cdot q_m$ synthesized user signals R^c and R^{bs} for all document clusters C_j ($j \in \{1, \ldots, q_m\}$). Note that we could not acknowledge what

Table 7. The Document Re-Ranking Performance in IRF, where Θ^{it} and $\Theta^{it,Sc}$ Indicate Fixed and Adaptable Combination Parameters, Respectively

Method[1]	NDCG@1	NDCG@3	NDCG@5	NDCG@10	MAP
$QE^{F_{\Theta^{it}}(R^c,R^p)}$	0.2842*	0.2952*	0.3124*	0.3690*	0.3708*
$QE^{F_{\Theta^{it}}(R^{bs},R^c,R^p)}$	0.3056*	0.3124*	0.3285*	0.3845*	0.3826*
$QE^{F_{\Theta^{it,Sc}}(R^c,R^p)}$	0.2948*	0.3024*	0.3191*	0.3747*	0.3744*
$QE^{F_{\Theta^{it,Sc}}(R^{bs},R^c,R^p)}$	**0.3126**	**0.3258**	**0.3505**	**0.4183**	**0.4061**

[1] $R^{\dagger}$ indicates relevance score based on signals †. bs, c, and p indicate brain signals, click signals, and pseudo-relevance signals, respectively. $*$ indicates a significant performance difference when comparing $QE^{F_{\Theta^{it}}(R^{bs},R^c,R^p)}$ to $QE^{F_{\Theta^{it,Sc}}(R^{bs},R^c,R^p)}$, with a significance level of $p < 1 \times 10^{-3}$.
Results in boldface denote the winning performance for each column.

Table 8. The Document Re-Ranking Performance in RRF, where Θ^{it} and $\Theta^{it,Sc}$ Indicate Fixed and Adaptable Combination Parameters, Respectively

Method[1]	NDCG@1	NDCG@3	NDCG@5	NDCG@10	MAP
$QE^{F_{\Theta^{it}}(R^{bs},R^c,R^p)}$	**0.6350**	0.6617	**0.7171**	0.7693	**0.8009**
$QE^{F_{\Theta^{it,Sc}}(R^{bs},R^c,R^p)}$	**0.6350**	**0.6622**	0.7170	**0.7694**	0.8004

[1] $R^{\dagger}$ indicates relevance score based on signals †. bs, c, and p indicate brain signals, click signals, and pseudo-relevance signals, respectively. The utilization of adaptable combination parameters $\Theta^{re,Sc}$ instead of fixed combination parameters Θ^{re} does not yield any significant differences.
Results in boldface denote the winning performance for each column.

subtopic a search engine user is interested in advance when they submit the query. Hence, the optimal combination parameters that have average best performances for all subtopics are selected by the proposed method.

5.4.6 Experimental Results. Table 7 presents the experimental results of IRF using adaptive RF signals combination methods and fixed combination methods. From Table 7, we observe that using adaptive RF combination methods is beneficial to IRF performance. This finding verifies our analyses and findings in Section 5.3. Furthermore, we observe that the performance improvement brought by brain signals differs depending on whether applying the adaptive RF signals combination method or not. Specifically, the additional improvement brought by brain signals is 1.5% ($QE^{F_{\theta^{it}}(R^{bs},R^c,R^p)}$ in comparison with $QE^{F_{\theta^{it}}(R^c,R^p)}$) in terms of NDCG@10 when using fixed combination parameters. On the other hand, the performance difference of $QE^{F_{\theta}^{it,Sc}(R^{bs},R^c,R^p)}$ and $QE^{F_{\theta}^{it,Sc}(R^c,R^p)}$ is 8.8% in terms of NDCG@10. This indicates that the adaptive RF signals combination method can further exploit the benefits brought by brain signals in the context of RF.

However, we observe that adopting the same adaptive RF signals combination method in RRF leads to no significant performance difference, as shown in Table 8. For example, the performance difference between $F_{\Theta}^{it,Sc}(R^{bs},R^c,R^p)$ and $F_{\Theta}^{it}(R^{bs},R^c,R^p)$ is not significant in terms of NDCG@10, where $\Theta^{it,Sc}$ is the adaptive combination parameter for RRF. This is due to the potential of the adaptive RF signals combination method is limited in RRF, as detailed in Section 5.3.1. The exploration on how to better utilize multiple RF signals in RRF is left as future work.

Moreover, the experimental results show the possibility of improving RF performance by adaptively adjusting the combination weight of different RF signals. However, the proposed algorithm requires document clustering for each query. In our experiment, the selected queries are broad and sometimes ambiguous, and the document clusters are usually related to the user's different search

intents. Therefore, whether this algorithm is suitable for other types of real-world queries remains unknown. The exploration of designing more general algorithms to improve RF performance is left as future work.

*Answer to **RQ4***. We devise an adaptive RF signals combination method to re-weight the importance of various RF signals (brain signals, click signals, and pseudo-relevance signals) according to the search scenarios. We observe significant performance improvement with the proposed method. This verifies the analyses in Section 5.3 and illustrates that brain signals can boost IRF performance with an improvement of 8.8% in terms of NDCG@10.

6 CONCLUSIONS AND DISCUSSIONS

6.1 Summary of Contributions

In this article, we propose a novel RF framework that combines pseudo-relevance signals, click signals, and brain signals for document re-ranking. Based on the proposed framework, we explore and verify the effectiveness of brain signals in the context of RF with different settings, that is, IRF and RRF. In addition to analyzing the overall performance improvement brought by brain signals, we also dive into several search scenarios where click signals are missing or biased. We observe that brain signals are more helpful to IRF at the beginning of the search process before any clicks are received. We also demonstrate that brain signals can be applied to identify "bad click" and improve the performance of RRF. Besides, we further analyze how the RF performance varies with the combination weights across different search scenarios. We demonstrate that the importance of brain signals and click signals could vary in different scenarios. Based on this observation, we then propose to adaptively combine different RF relevance signals based on search scenarios, which leads to additional performance improvement.

6.2 Discussions and Applications

With the development of neurological devices, several applications have emerged to improve the performance of interactive information systems with BCI, for example, personalized image editing [17], crowdsourcing [18, 22]. In IR, researchers also explore the possibility of establishing a pure BCI-based search system (which does not require mouse or keyboard). This opens up new avenues for understanding user intentions and enhancing search quality, while also highlighting potential challenges and concerns associated with brain-based RF.

6.2.1 Privacy Considerations. The potential of leveraging brain responses in information systems is accompanied by significant privacy concerns, particularly regarding the possible abuse of data gathered from BCIs. While the capabilities of anonymous technology in the field of brain signal processing are a cause for concern [10], recent advancements have tried to mitigate this issue by edge computing, where users have the right to decide whether to upload data to the cloud. While our research does not delve into the intricacies of these technologies, it provides key insights into the privacy landscape: On the one hand, our findings underscore that non-deep brain decoding models with rapid inferencing ability can yield substantial results, thus accentuating the viability of edge computing. Additionally, even without user-specific data for model training, cross-subject models can adequately support RF, particularly in cold-start situations. On the other hand, we observe that not every search scenario requires brain-based RF, but the importance of brain signals becomes paramount in specific search contexts. As such, users can opt to employ it selectively for tasks of higher difficulty.

6.2.2   EEG Availability. Currently, wearable EEG devices are applied in tandem with VR technologies [65] and in services catering to the differently-abled [12]. In IR-related research, functional information systems utilizing BCI have been developed. A notable instance is the work by Chen et al. [12], who crafted a fully operational BCI-integrated search system that eliminates the need for traditional input devices such as keyboard and mouse. The decrease in cost and increase in portability of BCIs have been recognized as significant factors when applying BCI to IR research and daily applications. Recently, Maiseli et al. [48] addresses the affordability and portability of BCIs, specifically mentioning the emergence of low-cost BCIs for everyday use, highlighting the progress towards making these technologies more accessible to the general public. This research elucidates the ongoing efforts to reduce the cost and enhance the portability of BCIs, which in turn, is expected to significantly broaden the spectrum of their application scenarios, including in the field of IR. What we add on top of existing BCI research is that we reveal another benefit of BCI for information systems, that is, RF, besides directly controlling information systems. The closed loop between the user and the system can be built since the system can better understand the user with RF model.

6.2.3   Comparison with Conventional User Signals. Since traditional interaction paradigms can only capture indirect user feedback, biases in these behaviors have been long-standing IR research problems. To address these challenges, we propose utilizing brain signals as a unique form of "explicit feedback". Concurrently, we identify specific search scenarios in which biases in clicks may adversely affect RF performance, highlighting search scenarios where brain signals can offer significant advantages.

6.2.4   Recommended Practices for EEG-Based Human Study. In our study, we aimed to implement settings that closely resemble real-world scenarios. For instance, instead of using random data splitting as done in prior research, we adopted a split-by-timepoint splitting to simulate the cold start situation that new users experience when they begin using our system. Furthermore, we conducted sensitivity analyses to demonstrate the impact of EEG pre-processing methods on relevance prediction performance. Our study provides reliable evidence for the practical use of EEG signal processing in IR scenarios.

6.2.5   Application Scenarios. **Interactive IR:** We reveal the possibility of establishing more interactive IR systems with brain signals by devising a novel RF framework. This gives insights into existing BCI-based search systems, in which users can not only search with their thoughts (without hand-based interactions) but also improve their search quality with a novel human-machine loop based on RF.

Human research for IR: It has been widely recognized that RF is not only a potentially useful technique for improving search quality but also an effective tool to investigate how people search [64]. The combined relevance score can not only be utilized for document re-ranking (IRF and RRF), but also help us understand how people actually interact with systems. Especially in search scenarios in which conventional RF signals may be biased, where brain signals are verified as a helpful substitute to infer relevance.

Human-Computer Interaction (HCI) applications: In addition to Web search scenarios, our RF framework can provide inspiration to several vertical applications related to IR. For example, a conversational chatbot may benefit from RF, especially in cases where users have difficulty expressing their thoughts in plain text but have ideas about the returned content [36]. Note that our proposed method for combining various RF signals is not restricted to combining brain signals

with click signals and pseudo-relevance signals in Web search scenarios. Hence this RF method can combine with other types of user signals and broader search applications and HCI products.

6.3 Limitations and Future Work

We evaluate the feasibility of combining brain signals and conventional feedback signals for two typical RF tasks, that is, IRF and RRF, for the first time. The experiments and analyses not only demonstrate that brain signals are feasible sources for RF, but also illustrate several search scenarios in which brain signals are more helpful to existing RF. Nonetheless, our research is not without its limitations, which may guide future work as follows:

6.3.1 Selection of Search Tasks. We derived our search tasks from TREC and iMine, characterized by their typically short and broad-topic queries. Given that RF does not universally benefit all queries and search scenarios [4], we prioritized these queries due to their heightened potential for RF advancements compared to more specific ones. This makes them a reasonable starting point to investigate the impact of brain signals on such standard queries. Future work may delve into a wider variety of query types and their synergy with brain signals in RF. Additionally, our approach extends beyond conventional query-based searches. The potential of our proposed RF framework also lies in its applicability to emerging search systems, like conversational systems, where traditional documents transition into diverse system response formats.

6.3.2 User Study Settings. First, our user study adopted a relatively strict design to mitigate potential confounders affecting brain signal collection. For example, complex search behaviors in real life such as revisit [73], and comparison [84] are not considered in our study to avoid uncontrolled bias. This constraint is not unique to our approach but is inherent in any behavioral signals [5] used for search performance evaluation. Nevertheless, our work pioneers the integration of brain signals in RF within web search contexts. Contrasted with existing NeuraSearch literature [53, 61], our study employs more authentic queries and Web pages sourced from prevalent commercial search engines. The challenge of navigating intricate scenarios with unfettered user behaviors remains a prospect for subsequent studies.

6.3.3 Comparison with More User Signals. In IR research, apart from click signals and brain signals, there are also user signals such as eye-tracking and mouse movements. However, existing research has shown that eye-tracking and mouse movements are not accurate enough since they are also indirect probes of user signals [49]. The main difference between brain signals and conventional signals is that brain signals provide direct feedback on a person's consciousness and thoughts. The only concern lies in the noise introduced during brain signal collection and decoding, which can be further mitigated through improved equipment and data quality. In our study, we only compare brain signals with pseudo-relevance signals and click signals since they are the most commonly used signals in RF. However, the idea of integrating eye movement data and more modalities in a future investigation is intriguing and holds merit.

REFERENCES

[1] IJsbrand Jan Aalbersberg. 1992. Incremental relevance feedback. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 11–22.

[2] Stephen Akuma. 2022. Eye gaze relevance feedback indicators for information retrieval. *International Journal of Intelligent Systems and Applications* 14, 1 (2022), 57–65.

[3] 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 *Proceedings of the 38th International Acm Sigir Conference on Research and Development in Information Retrieval*. 719–722.

[4] Hiteshwar Kumar Azad and Akshay Deepak. 2019. Query expansion techniques for information retrieval: A survey. *Information Processing & Management* 56, 5 (2019), 1698–1735.

[5] Leif Azzopardi. 2021. Cognitive biases in search: A review and reflection of cognitive biases in information retrieval. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 27–37.

[6] Aayush Bhardwaj, Ankit Gupta, Pallav Jain, Asha Rani, and Jyoti Yadav. 2015. Classification of human emotions from EEG signals using SVM and LDA classifiers. In *Proceedings of the 2015 2nd International Conference on Signal Processing and Integrated Networks* . IEEE, 180–185.

[7] Keping Bi, Qingyao Ai, and W Bruce Croft. 2019. Revisiting iterative relevance feedback for document and passage retrieval. *SIGIR Workshop on Conversational Interaction Systems (WCIS'19)*.

[8] Keping Bi, Qingyao Ai, and W. Bruce Croft. 2019. Iterative relevance feedback for answer passage retrieval with passage-level semantic match. In *Proceedings of the Advances in Information Retrieval: 41st European Conference on IR Research, Proceedings, Part I 41*. Springer, 558–572.

[9] Keping Bi, Qingyao Ai, Yongfeng Zhang, and W. Bruce Croft. 2019. Conversational product search based on negative feedback. In *Proceedings of the 28th Acm International Conference on Information and Knowledge Management*. 359–368.

[10] Amir Jalaly Bidgoly, Hamed Jalaly Bidgoly, and Zeynab Arezoumand. 2022. Towards a universal and privacy preserving EEG-based authentication system. *Scientific Reports* 12, 1 (2022), 2531.

[11] Ethem F. Can, W. Bruce Croft, and R. Manmatha. 2014. Incorporating query-specific feedback into learning-to-rank models. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1035–1038.

[12] Xuesong Chen, Ziyi Ye, Xiaohui Xie, Yiqun Liu, Xiaorong Gao, Weihang Su, Shuqi Zhu, Yike Sun, Min Zhang, and Shaoping Ma. 2022. Web search via an efficient and effective brain-machine interface. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*. 1569–1572.

[13] Aleksandr Chuklin, Ilya Markov, and Maarten De Rijke. 2022. *Click Models for Web Search*. Springer Nature.

[14] Charles LA Clarke, Nick Craswell, and Ian Soboroff. 2009. Overview of the TREC 2009 Web track. In *Proceedings of the 18th Text REtrieval Conference*, Vol. 9. 20–29.

[15] Mark Claypool, Phong Le, Makoto Wased, and David Brown. 2001. Implicit interest indicators. In *Proceedings of the 6th International Conference on Intelligent User Interfaces*. 33–40.

[16] Paul Clough and Mark Sanderson. 2013. Evaluating the performance of information retrieval systems using test collections. *Information Research* 18, 2 (2013), 18–2.

[17] Keith M. Davis, Carlos de la Torre-Ortiz, and Tuukka Ruotsalo. 2022. Brain-supervised image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18480–18489.

[18] Keith M. Davis III, Lauri Kangassalo, Michiel Spapé, and Tuukka Ruotsalo. 2020. Brainsourcing: Crowdsourcing recognition tasks via collaborative brain-computer interfacing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[19] Keith M. Davis III, Michiel Spapé, and Tuukka Ruotsalo. 2021. Collaborative filtering with preferences inferred from brain signals. In *Proceedings of the Web Conference 2021*. 602–611.

[20] Daniel de Queiroz Cavalcanti, Felipe Melo, Thiago Silva, Matheus Falcão, Matheus Cavalcanti, and Valdecir Becker. 2023. Research on brain-computer interfaces in the entertainment field. In *Proceedings of the International Conference on Human-Computer Interaction*. Springer, 404–415.

[21] Abel Desoto, Joshua Dodd, Micheal Babinec, and Kiran George. 2023. Utilization of EEG and fNIRS To determine neural alignment in educational applications. In *Proceedings of the 2023 IEEE World AI IoT Congress*. IEEE, 0155–0157.

[22] Suzanne Dikker, Georgios Michalareas, Matthias Oostrik, Amalia Serafimaki, Hasibe Melda Kahraman, Marijn E. Struiksma, and David Poeppel. 2021. Crowdsourcing neuroscience: Inter-brain coupling during face-to-face interactions outside the laboratory. *NeuroImage* 227, 1 (2021), 117436.

[23] Yeshi Dolma, Raunak Kalani, Astha Agrawal, and Saurav Basu. 2021. Improving bounce rate prediction for rare queries by leveraging landing page signals. In *Companion Proceedings of the Web Conference 2021*. 1–6.

[24] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang L]u. 2013. Differential entropy feature for EEG-based emotion classification. In *Proceedings of the 2013 6th International IEEE/EMBS Conference on Neural Engineering*. IEEE, 81–84.

[25] Manuel J. A. Eugster, Tuukka Ruotsalo, Michiel M. Spapé, Oswald Barral, Niklas Ravaja, Giulio Jacucci, and Samuel Kaski. 2016. Natural brain-information interfaces: Recommending information by relevance inferred from human brain signals. *Scientific Reports* 6, 1 (2016), 38580.

[26] Manuel J. A. Eugster, Tuukka Ruotsalo, Michiel M. Spapé, Ilkka Kosunen, Oswald Barral, Niklas Ravaja, Giulio Jacucci, and Samuel Kaski. 2014. Predicting term-relevance from brain signals. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. 425–434.

[27] Jan-Eike Golenia, Markus A. Wenzel, Mihail Bogojeski, and Benjamin Blankertz. 2018. Implicit relevance feedback from electroencephalography and eye tracking in image search. *Journal of Neural Engineering* 15, 2 (2018), 026002.

[28] Jacek Gwizdka, Rahilsadat Hosseini, Michael Cole, and Shouyi Wang. 2017. Temporal dynamics of eye-tracking and EEG during reading and relevance decisions. *Journal of the Association for Information Science and Technology* 68, 10 (2017), 2299–2312.

[29] Donna Harman. 1992. Relevance feedback revisited. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1–10.

[30] Hilde M. Huizenga, Jan C. De Munck, Lourens J. Waldorp, and Raoul P. P. P. Grasman. 2002. Spatiotemporal EEG/MEG source analysis based on a parametric noise covariance model. *IEEE Transactions on Biomedical Engineering* 49, 6 (2002), 533–539.

[31] Aapo Hyvärinen. 1997. New approximations of differential entropy for independent component analysis and projection pursuit. *Advances in Neural Information Processing Systems* 10 (1997), 273–279.

[32] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20, 4 (2002), 422–446.

[33] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*. 781–789.

[34] Jukka-Pekka Kauppi, Melih Kandemir, Veli-Matti Saarinen, Lotta Hirvenkari, Lauri Parkkonen, Arto Klami, Riitta Hari, and Samuel Kaski. 2015. Towards brain-activity-controlled information retrieval: Decoding image relevance from MEG signals. *NeuroImage* 112 (2015), 288–298.

[35] Aleksandra Kawala-Sterniuk, Natalia Browarska, Amir Al-Bakri, Mariusz Pelc, Jaroslaw Zygarlicki, Michaela Sidikova, Radek Martinek, and Edward Jacek Gorzelanczyk. 2021. Summary of over fifty years with brain-computer interfacesâĂŤa review. *Brain Sciences* 11, 1 (2021), 43.

[36] Johannes Kiesel, Damiano Spina, Henning Wachsmuth, and Benno Stein. 2021. The meant, the said, and the understood: Conversational argument search and cognitive biases. In *Proceedings of the 3rd Conference on Conversational User Interfaces*. 1–5.

[37] Sander Koelstra, Ch]ristian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing* 3, 1 (2011), 18–31.

[38] Varun Kohli, Utkarsh Tripathi, Vinay Chamola, Bijay Kumar Rout, and Salil S. Kanhere. 2022. A review on Virtual Reality and Augmented Reality use-cases of Brain Computer Interface based applications for smart cities. *Microprocessors and Microsystems* 88 (2022), 104392.

[39] Zirui Lan, Olga Sourina, Lipo Wang, Reinhold Scherer, and Gernot R. Müller-Putz. 2018. Domain adaptation techniques for EEG-based emotion recognition: A comparative study on two public datasets. *IEEE Transactions on Cognitive and Developmental Systems* 11, 1 (2018), 85–94.

[40] Victor Lavrenko and W. Bruce Croft. 2017. Relevance-based language models. *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, 260–267.

[41] Hang Li, Ahmed Mourad, Bevan Koopman, and Guido Zuccon. 2022. How does feedback signal quality impact effectiveness of pseudo relevance feedback for passage retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2154–2158.

[42] Jane Li, Scott Huffman, and Akihito Tokuda. 2009. Good abandonment in mobile and PC internet search. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 43–50.

[43] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2021. Multistage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting. *ACM Transactions on Information Systems* 39, 4 (2021), 1–29.

[44] Yiqun Liu, Ruihua Song, Min Zhang, Zhicheng Dou, Takehiro Yamamoto, Makoto P. Kato, Hiroaki Ohshima, and Ke Zhou. 2014. Overview of the NTCIR-11 IMine task. In *Proceedings of the NTCIR Conference on Evaluation of Information Access Technologies*.

[45] 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 *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*. 849–858.

[46] Hongyu Lu, Min Zhang, and Shaoping Ma. 2018. Between clicks and satisfaction: Study on multi-phase user preferences and satisfaction for online news reading. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 435–444.

[47] Yuanhua Lv and ChengXiang Zhai. 2009. Adaptive relevance feedback in information retrieval. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. 255–264.

[48] Baraka Maiseli, Abdi T. Abdalla, Libe V. Massawe, Mercy Mbise, Khadija Mkocha, Nassor Ally Nassor, Moses Ismail, James Michael, and Samwel Kimambo. 2023. Brain–computer interface: Trend, challenges, and threats. *Brain Informatics* 10, 1 (2023), 20.

[49] Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2014. Estimating credibility of user clicks with mouse movement and eye-tracking information. In *Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 263–274.

[50] Ali Montazeralghaem, Hamed Zamani, and James Allan. 2020. A reinforcement learning framework for relevance feedback. In *Proceedings of the 43rd International Acm Sigir Conference on Research and Development in Information Retrieval.* 59–68.

[51] Masahiro Morita and Yoichi Shinoda. 1994. Information filtering based on user behavior analysis and best match text retrieval. In *SIGIR'94: Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval.* Springer, 272–281.

[52] Yashar Moshfeghi and Joemon M. Jose. 2013. On cognition, emotion, and interaction aspects of search tasks with different search intentions. In *Proceedings of the 22nd International Conference on World Wide Web.* 931–942.

[53] Yashar Moshfeghi, Luisa R. Pinto, Frank E. Pollick, and Joemon M. Jose. 2013. Understanding relevance: An fMRI study. In *Proceedings of the Advances in Information Retrieval: 35th European Conference on IR Research. Proceedings 35.* Springer, 14–25.

[54] Yashar Moshfeghi and Frank E. Pollick. 2018. Search process as transitions between neural states. In *Proceedings of the 2018 World Wide Web Conference.* 1683–1692.

[55] Yashar Moshfeghi, Peter Triantafillou, and Frank E. Pollick. 2016. Understanding information need: An fMRI study. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 335–344.

[56] Javed Mostafa and Jacek Gwizdka. 2016. Deepening the role of the user: Neuro-physiological evidence as a basis for studying and improving search. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval.* 63–70.

[57] Liad Mudrik, Dominique Lamy, and Leon Y. Deouell. 2010. ERP evidence for context congruity effects during simultaneous object–scene processing. *Neuropsychologia* 48, 2 (2010), 507–517.

[58] Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. Rethink training of BERT rerankers in multi-stage retrieval pipeline. *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, Springer, 280–286.

[59] Mateus Pereira, Elham Etemad, and Fernando Paulovich. 2020. Iterative learning to rank from explicit relevance feedback. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing.* 698–705.

[60] Gert Pfurtscheller and F. H. Lopes Da Silva. 1999. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology* 110, 11 (1999), 1842–1857.

[61] Zuzana Pinkosova, William J. McGeown, and Yashar Moshfeghi. 2020. The cortical activity of graded relevance. In *Proceedings of the 43rd International ACM Sigir Conference on Research and Development in Information Retrieval.* 299–308.

[62] Stephen Robertson, Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[63] Joseph John Rocchio Jr. 1971. Relevance feedback in information retrieval. *The SMART retrieval system: Experiments in automatic document processing* (1971).

[64] Ian Ruthven and Mounia Lalmas. 2003. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review* 18, 2 (2003), 95–145.

[65] Jan-Philipp Tauscher, Fabian Wolf Schottky, Steve Grogorick, Paul Maximilian Bittner, Maryam Mustafa, and Marcus Magnor. 2019. Immersive EEG: Evaluating electroencephalography in virtual reality. In *Proceedings of the 2019 IEEE Conference on Virtual Reality and 3D User Interfaces.* IEEE, 1794–1800.

[66] Xiao Wang, Craig Macdonald, Nicola Tonellotto, and Iadh Ounis. 2021. Pseudo-relevance feedback for multiple representation dense retrieval. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval.* 297–306.

[67] Xiao Wang, Craig Macdonald, Nicola Tonellotto, and Iadh Ounis. 2023. ColBERT-PRF: Semantic pseudo-relevance feedback for dense passage and document retrieval. *ACM Transactions on the Web* 17, 1 (2023), 1–39.

[68] Ryen W. White, Joemon M. Jose, and Ian Ruthven. 2006. An implicit feedback approach for interactive information retrieval. *Information Processing & Management* 42, 1 (2006), 166–190.

[69] Ryen W. White, Ian Ruthven, and Joemon M. Jose. 2002. The use of implicit evidence for relevance feedback in web retrieval. In *Proceedings of the Advances in Information Retrieval: 24th BCS-IRSG European Colloquium on IR Research Proceedings 24.* Springer, 93–109.

[70] Irene Winkler, Stefan Haufe, and Michael Tangermann. 2011. Automatic classification of artifactual ICA-components for artifact removal in EEG signals. *Behavioral and Brain Functions* 7 (2011), 1–15.

[71] Dan Wu, Jing Dong, Li Shi, Chunxiang Liu, and Jiangyun Ding. 2020. Credibility assessment of good abandonment results in mobile search. *Information Processing & Management* 57, 6 (2020), 102350.

[72] Xiaohui Xie, Qian Dong, Bingning Wang, Feiyang Lv, Ting Yao, Weinan Gan, Zhijing Wu, Xiangsheng Li, Haitao Li, Yiqun Liu, and Jin Ma. 2023. T2Ranking: A large-scale chinese benchmark for passage ranking. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'23)*, Association for Computing Machinery, New York, NY, 2681–2690. https://doi.org/10.1145/3539618.3591874

[73] Danqing Xu, Yiqun Liu, Min Zhang, Shaoping Ma, and Liyun Ru. 2012. Incorporating revisiting behaviors into click models. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*. 303–312.

[74] Haopei Yang, Geoffrey Laforge, Bobby Stojanoski, Emily S. Nichols, Ken McRae, and Stefan Köhler. 2019. Late positive complex in event-related potentials tracks memory signals when they are decision relevant. *Scientific Reports* 9, 1 (2019), 9469.

[75] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 72–77.

[76] Xiaolan Yang, Mei Gao, Jinchuan Shi, Hang Ye, and Shu Chen. 2017. Modulating the activity of the DLPFC and OFC has distinct effects on risk and ambiguity decision-making: A tDCS study. *Frontiers in Psychology* 8 (2017), 1417.

[77] Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: BERT and beyond. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 1154–1156.

[78] Ziyi Ye, Xiaohui Xie, Yiqun Liu, Zhihong Wang, Xuesong Chen, Min Zhang, and Shaoping Ma. 2022. Brain topography adaptive network for satisfaction modeling in interactive information access system. In *Proceedings of the 30th ACM International Conference on Multimedia*. 90–100.

[79] Ziyi Ye, Xiaohui Xie, Yiqun Liu, Zhihong Wang, Xuancheng Li, Jiaji Li, Xuesong Chen, Min Zhang, and Shaoping Ma. 2022. Why don't you click: Understanding non-click results in web search with brain signals. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 633–645.

[80] Hongzhi Yin, Bin Cui, Ling Chen, Zhiting Hu, and Zi Huang. 2014. A temporal context-aware model for user behavior modeling in social media systems. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. 1543–1554.

[81] HongChien Yu, Chenyan Xiong, and Jamie Callan. 2021. Improving query representations for dense retrieval with pseudo relevance feedback. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3592–3596.

[82] Fan Zhang, Yiqun Liu, Jiaxin Mao, Min Zhang, and Shaoping Ma. 2020. User behavior modeling for web search evaluation. *AI Open* 1 (2020), 40–56.

[83] Junqi Zhang, Yiqun Liu, Shaoping Ma, and Qi Tian. 2018. Relevance estimation with multiple information sources on search engine result pages. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 627–636.

[84] Ruizhe Zhang, Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2021. Constructing a comparison-based click model for web search. In *Proceedings of the Web Conference 2021*. 270–283.

[85] Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. 2020. BERT-QE: Contextualized query expansion for document re-ranking. *Findings of the Association for Computational Linguistics (EMNLP'20)*.

[86] Peixiang Zhong, Di Wang, and Chunyan Miao. 2020. EEG-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing* 13, 3 (2020), 1290–1301.