Why Don't You Click: Understanding Non-Click Results in Web Search with Brain Signals

Ziyi Ye
yeziyi1998@gmail.com
BNRist,DCST,Tsinghua University
Beijing, China

Xiaohui Xie
xiexh_thu@163.com
BNRist,DCST,Tsinghua University
Beijing, China

Yiqun Liu*
yiqunliu@tsinghua.edu.cn
BNRist,DCST,Tsinghua University
Beijing, China

Zhihong Wang
wangzhh629@mail.tsinghua.edu.cn
BNRist,DCST,Tsinghua University
Beijing, China

Xuancheng Li
lixuanch18@mails.tsinghua.edu.cn
BNRist,DCST,Tsinghua University
Beijing, China

Jiaji Li
jiajili@link.cuhk.edu.cn
SDS, The Chinese University of Hong Kong, Shenzhen, China

Xuesong Chen
chenxuesong1128@163.com
BNRist,DCST,Tsinghua University
Beijing, China

Min Zhang
z-m@tsinghua.edu.cn
BNRist,DCST,Tsinghua University
Beijing, China

Shaoping Ma
msp@tsinghua.edu.cn
BNRist,DCST,Tsinghua University
Beijing, China

ABSTRACT

Web search heavily relies on click-through behavior as an essential feedback signal for performance evaluation and improvement. Traditionally, click is usually treated as a positive implicit feedback signal of relevance or usefulness, while non-click is regarded as a signal of irrelevance or uselessness. However, there are many cases where users satisfy their information need with the contents shown on the Search Engine Result Page (SERP). This raises the problem of measuring the usefulness of non-click results and modeling user satisfaction in such circumstances.

For a long period, understanding non-click results is challenging owing to the lack of user interactions. In recent years, the rapid development of neuroimaging technologies constitutes a paradigm shift in various industries, e.g., search, entertainment, and education. Therefore, we benefit from these technologies and apply them to bridge the gap between the human mind and the external search system in non-click situations. To this end, we analyze the differences in brain signals between the examination of non-click search results in different usefulness levels. Inspired by these differences, we conduct supervised learning tasks to estimate the usefulness of non-click results with brain signals and conventional information (i.e., content and context factors). Furthermore, we devise two re-ranking methods, i.e., a Personalized Method (PM) and a Generalized Intent modeling Method (GIM), for search result re-ranking with the estimated usefulness. Results show that it is feasible to utilize brain signals to improve usefulness estimation performance and enhance human-computer interactions by search result re-ranking.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Users and interactive retrieval**.

KEYWORDS

Zero-click Search, Good Abandonment, Click Necessity, Usefulness, Brain Signals, EEG

ACM Reference Format:
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 (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3477495.3532082

1 INTRODUCTION

The Information Retrieval (IR) community has a long tradition of using click-through behavior as vital user feedback for search evaluation [16] and relevance modeling [17, 34]. These researches usually consider click as a positive signal for relevance or usefulness and non-click (especially non-click after examination) as negative. However, search results returned by the current search engines are far more informative than "ten blue links", aiming to satisfy the user's information need without any click on the Search Engine Result Page (SERP). Figure 1 presents examples of three real-world search results, of which two are non-click.

With the advancement of search engines, it is prevalent to find non-click results useful. Previous literature has investigated an extreme case called "Zero-click" search [1], where users do not click on any results in a search session. As reported, "Zero-click" searches on

*Yiqun Liu is the corresponding author.

SIGIR '22, July 11–15, 2022, Madrid, Spain.

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3532082

[1] https://www.searchmetrics.com/glossary/zero-click-searches/

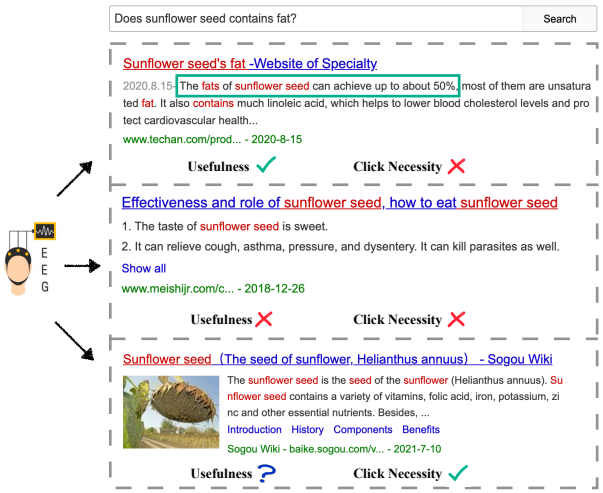


Figure 1: Examples of non-click search results. The first result is helpful to satisfy the user's information need with its snippets and unnecessary to click.

Google have risen to nearly 65% in 2020 [2]. Therefore, understanding why the user does not click a search result or even abandon a SERP becomes a vital challenge and attracts much attention. Existing efforts focus on detecting "good abandonment" (i.e., user satisfy their Information Need (IN) without clicks) with SERP content [42] and search context [7, 42]. However, these works only involve a special case where all results are non-click ones and analyze the non-click behavior at a coarse-grained level, i.e., SERP-level. Besides, their prediction performances are limited by the lack of user interactions as implicit feedback, such as clicks on the SERP and interactions on the landing pages. Thus, there is still room for improving the performance if more feedback signals can be acquired.

Recently, the development of brain–computer interfaces (BCIs) makes it feasible to collect user feedback for non-click results and "Zero-click" scenarios. As BCI devices become low-cost (hundreds of dollars) and portable [3], BCI constitutes a paradigm shift in human–machine interaction and reshapes human's life in many domains, such as game playing [41] and image recommendation [5]. In IR domain, Liu et al. [24] suggest a revolution of search interface with portable BCI and Chen et al. [4] verify its feasibility by building a practical BCI controlled search system. In addition to utilizing brain signals to control search interface, decoding brain signals into user feedback for search performance improvement is another benefit of BCI-enhanced search. However, to what extent brain signals could benefit search performance, especially for non-click results that suffer from the lack of user feedback, remains an open problem.

In this paper, we delve into the phenomenon of non-click behavior and explore the effectiveness of usefulness estimation with different information sources (i.e., content, context, and brain signals, detailed in Section 5.1.1). The following research questions are raised:

- **RQ1**: How are brain signals in the non-click behaviors associated with the usefulness of search results?
- **RQ2**: To what extent can we estimate the usefulness of non-click results with additional information sources of brain signals?
- **RQ3**: Can we improve the performance of search result re-ranking with the estimated usefulness?

To shed light on these research questions, we conduct a lab-based user study to investigate non-click behavior. Participants are required to perform search tasks while an electroencephalogram (EEG) device is applied to collect their brain activities. The exploratory analyses indicate that the EEG band power is correlated with search result usefulness, particularly in the brain regions of left temporal, frontal, and occipital. This finding illustrates the possibility of utilizing brain signals for the usefulness estimation of non-click results.

To verify the effectiveness of brain signals in the usefulness estimation task, we conduct extensive experiments with user-independent and task-independent protocols. Experimental results demonstrate that models with brain signals can obtain a significant improvement of 6.8% (user-independent) and 11.9% (task-independent) in terms of AUC compared to the conventional usefulness estimation models based on content and context factors [26]. Furthermore, we propose two re-ranking methods, a Personalized Method (PM) and a Generalized Intent modeling Method (GM), for search result re-ranking with the estimated usefulness. By virtue of brain signals, the search result re-ranking task obtains a performance improvement of 17.0% and 20.6% in terms of $NDCG@1$ for PM and GIM, respectively. These experimental results illustrate that brain signals are valuable feedback during non-click search result examination. Furthermore, the findings also demonstrate the benefits of constructing a proactive search system with real-time BCI in the foreseeable future.

2 RELATED WORK

2.1 Zero-click Search

"Zero-click" refers to the situation that the SERP successfully and entirely satisfies the IN, without the necessity to click on a search result. Recently, commercial search engines have been attempting to improve user experience by extracting high-quality snippets or creating enhanced search results so that a user can pay as little effort (including click) as possible to access the IN. Therefore, "Zero-click" search plays an important role in real-world IR and attracts much attention.

To understand the non-click behaviors in Web search, recent researches have concentrated on "good abandonment", which indicates the user's IN is successfully realized with no need to click on a result or refine the query. For instance, Li et al. [21] approximate the prevalence of good abandonment in desktop and mobile search logs and find that a large amount of abandonment behavior is good abandonment, especially in mobile search. Additionally, some researchers detect and predict good abandonment in desktop [7] and mobile [42, 43], with the help of page content and user interactions.

However, these studies exclude the understanding of fine-grained usefulness for each result, which is more beneficial than page-level satisfaction for search evaluation. To unravel the usefulness of

[2] https://sparktoro.com/blog/in-2020-two-thirds-of-google-searches-ended-without-a-click/
[3] https://the-unwinder.com/reviews/best-eeg-headset/

non-click results and deal with the lack of user interactions in this scenario, we leverage the brain signals as user feedback and demonstrate its effectiveness. To our best knowledge, our work reveals the difference in brain activities while examining search results with different usefulness for the first time. We believe that our paradigm can extend to other situations that lack user interactions.

2.2 Usefulness of Search Result

In the user-centric evaluation, usefulness is a significant concept. Dislike relevance, which is often annotated by external assessors, *"usefulness represents users' opinions about whether search results can meet their INs"* [39]. Mao et al. [27] find that there exist many cases where high relevance doesn't mean the document is useful. And they reveal that usefulness has a higher correlation with user satisfaction than relevance. With such findings, they further propose models for usefulness judgment prediction in desktop search scenarios [26] and mobile search scenarios [25].

With the emergence of "Zero click" search, understanding the usefulness judgment of non-click results is vital. One of the challenges is the interactions on the landing page, which contain valuable feedback such as dwell time and mouse movement, are absent for non-click results. Therefore, we collect brain signals during the examination of non-click results to uncover this problem. In addition, to verify the effectiveness of collected brain signals, we evaluate and compare the usefulness estimation model based on brain signals and conventional features proposed by Mao et al. [26].

2.3 BCI for IR

There is an increasing number of literature that applies neurological devices to IR research. On the one hand, several studies investigate the cognitive components related to IR from a neuroscience perspective. For example, Moshfeghi et al. [28, 29, 30] and Pinkosova et al. [32] conduct a series of studies using brain signals to unravel the nature of a set of core notions, such as relevance and IN. They demonstrate the distributed network of brain regions associated with these concepts and related IR tasks. Insightful findings are obtained, such as (1) IN reflects a neural mechanism to acquire external information sources, and (2) relevance is a graded phenomenon in the human brain.

On the other hand, recent years have witnessed some researches utilize brain signals to infer relevance. For instance, Gwizdka et al. [10] conduct extensive studies to judge text relevance using EEG or in combination with eye movements. They show that models using EEG features can achieve an improvement of 20% in terms of AUC compared to that of an untrained model. And their work is further extended into classifying topical relevance of visual shots with EEG algorithm [18]. Nevertheless, these studies are not carried out in the search scenario, which includes the interactions on SERP and examination on the landing page. Whether it is possible to improve search performance with BCI remains unknown.

Recently, researchers suggest the possibility of using BCI as a new interface for search [24]. Inspired by the latest BCI technology, Chen et al. [4] design the first ready-to-use BCI-based search system with Steady-State Visual Evoked Potentials (SSVEP). It can help scenarios in which hand-based interactions are infeasible, e.g., virtual reality games and users with severe neuromuscular disorders.

Going one step further, we suggest that brain signals can not only control the search system but also benefit the search experience with real-time feedback, especially for non-click results that suffer from the lack of user interactions. What we add on top of these works is that we demonstrate the effectiveness of brain signals for the usefulness estimation of non-click results in search scenarios. Our work suggests that a BCI-based search system can provide additional user feedback for performance improvement and evaluation. We believe that our experimental results can verify the benefits of BCI-based search in the foreseeable future.

3 DATA COLLECTION

In this section, we introduce the design of our user study and the collected dataset [4].

3.1 User study tasks

We first select 150 queries from the SRR (Search Result Relevance) dataset [45] for our user study. We use this dataset for two main reasons: (1) It contains a large number of real-life query logs, screenshots of search results, and landing pages. Each query has ten corresponding results. (2) It provides human annotations of result type according to presentation styles. To ensure that a query is understandable and has a greater probability to cause non-click behaviors, 90 queries are sampled based on these criteria: (1) A sampled query should have a clear and unambiguous description and have a straightforward information need. (2) Among the ten corresponding results, at least four results are unnecessary to click. The click necessity is annotated by 15 external assessors. For each result, click necessity (binary) is judged by at least three different assessors, and the majority vote decides whether it is necessary to click. After the selection, we generate a task description manually for each task and collect the corresponding search results in the dataset for our study. Figure 1 presents an example query "Does sunflower seed contains fat?" and some of its corresponding search results (translated from Chinese). Note that we consider search tasks with straightforward information need in our user study, it's also interesting to explore the exploratory search scenarios in the furture.

3.2 Participants

We recruit 18 college students aged from 19 to 26 (M [5] = 21.56, SD [6] = 1.82). The number of participants is analogous to previous EEG-based studies (e.g., 15 in [15] and 20 in [1]) and the estimated sample size for the factor analysis in Section 4.1 is 18 (statistical power=0.8, α=0.05). There are ten males and eight females who mainly major in computer science, physics, arts, and engineering. All the participants are acquainted with the usage of search engines, and all of them report using search engines daily or once in two days. The whole task takes about two hours to complete: 50 minutes for preparation and rest, 60 minutes for the main task, and 10 minutes for the questionnaire procedure. And each participant would gain $30 after completing all the tasks.

[4] The data and code is publicly available in http://www.thuir.cn/Search_Brainwave/.
[5] Mean value.
[6] Standard deviation.

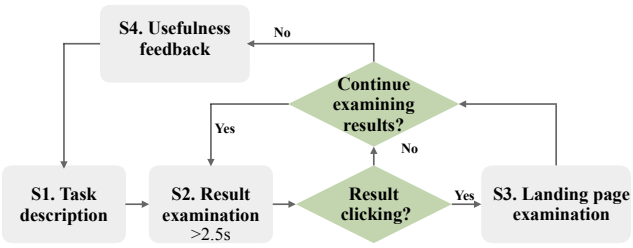


Figure 2: The procedure of a search task. If participants choose not to click in S2, the result is a non-click one.

3.3 Procedure

This user study adheres to the ethical procedures which is approved by the ethics committee of the School of Psychology at Tsinghua University. In the beginning, participants fill in an entry questionnaire to report demographic information and sign an informed consent about security and privacy protection. Then they read user study instructions about the procedure of each search task during the user study. Before entering the main step, participants undergo a training step with two search tasks to ensure they are familiar with the procedure. Each participant is instructed to complete search tasks on a website developed with Django. In the main step, the participants are supposed to seriously accomplish the search tasks as many as possible in 60 minutes. They are allowed to rest between tasks while the rest time would not be included in the time limit.

Figure 2 illustrates the procedure of each search task in the main step. For simplicity, we denote a *search result* is a screenshot on the SERP (Figure 1 gives three examples) and its corresponding *landing page* is the standalone web page after the user click the link corresponding to the search result. The search tasks follow the same order of steps, i.e., S1 to S4:

(S1) Participants view a task description randomly selected from the dataset. Once they fully understand the question, they can press a button and enter the second step.

(S2) A fixation cross is presented for 1.5 seconds on the screen center to capture participants' attention and indicate the location of the forthcoming result. Then a search result (Figure 1 gives three examples) is displayed, lasting for 2.5 seconds. This procedure, following the previous works [30, 44], ensures that brain activity related to the motor response of moving the cursor and clicking the button would not be contained during the 2.5 seconds. And we would use brain signals recorded in this time interval for further analysis and experiments. After that, three response choices, i.e., "skip", "click", and "end the search", will be presented above the search result. If the user chooses to skip the search result or end the search, the search result will be a non-click one.

(S3) If participants choose "click" in (S2), the landing page of the corresponding search result will be presented. After examining the lading page, the participant can either end the search or continue to examine the next result in this step.

(S4) Once the participant is convinced of the answer to the search task, they can end the search in (S2) or (S3). Then they are presented with an end-mark page. On this page, they are required to give the answer via voice input and report their perceived difficulty (five-point Likert scale) to the search task and usefulness feedback (four-point Likert scale) to each result. The participants are informed that the voice input would be examined to ensure that they have carefully accomplished the search tasks.

We randomize the tasks' order for each participant and display each search result within a task in a randomized sequence. A pilot study, which involved four additional users, is conducted ahead to adjust the settings, including the display time of fixation cross and result, amount of training tasks, etc. Note that in our experimental paradigm, the search result is displayed one by one. We apply this paradigm to collect brain signals and behavior responses for each specific result and leave the investigation on the whole page as future work.

3.4 Preprocessing of EEG data

EEG data commonly contains noise sources related to power line noise, eye blinks, body movement, etc., which need to be pre-processed with standard procedures for further analysis. The standard procedures include: re-referencing to averaged mastoids, baseline correlation, low-pass of 50Hz and high-pass of 0.5Hz filtering, artifacts removal (with a parametric noise covariance model [12]), and down-sampling to 500 Hz. Afterward, interested epochs (brief EEG segment, 2,500 ms in our experimental settings) are extracted, and baseline correlation is applied again using the pre-stimulus period 0-1500 ms.

3.5 Apparatus

Our study uses a desktop computer that has a 27-inch monitor with a resolution of 2,560×1440 and Google Chrome browser. A Scan NuAmps Express system (Compumedics Ltd., VIC, Australia) and a 64-channel Quik-Cap (Compumedical NeuroScan) are deployed to capture the participants' EEG data. All the EEG channels are placed based on the International 10–20 system. The impedance of the channels is calibrated under 10 $k\Omega$ in the preparation step, and the sampling rate is set at 1,000 Hz.

3.6 Statistics of the Collected Data

The collected dataset consists of 1252 interactions on 90 search tasks, and participants examine 3.61(SD=2.24) search results for each task on average. One participant averagely accomplishes 69.56 (SD=12.23) tasks and examines 250.78 (SD=56.53) search results. Table 1 presents the participants' responses (click and non-click) across usefulness levels ranging from 1 to 4. We can observe that about 85.9% of search results are non-clicked, among which 46.8% are "not useful at all" (usefulness=1), followed by "very useful" (usefulness=4), while fewer in "fairly useful" (usefulness=3) and

Table 1: The average number of participants' responses across usefulness levels.

Response	Usefulness			
	1	2	3	4
#Click	14.0(±14)	7.4(±7)	10.2(±11)	12.8(±13)
#Non-click	101.2(±53)	30.8(±21)	31.5(±20)	52.5(±16)

Table 2: Statistical significant differences in EEG spectral powers among various coarse-grained brain regions. (*/, ·) indicates the ANOVA test is significant at the $p < 0.05/0.01$ level. (·, $\uparrow^{*/**}$ /$\downarrow^{*/**}$) indicates the post-hoc test suggest the spectral powers in "not useful at all" group in higher/lower than that in the "very useful" group at the $p < 0.05/0.01$ level.**

Brain region	Bands (ANOVA test, post-hoc test)
Pre-frontal	δ (*, $\uparrow$**), θ (**, $\downarrow$**), β (*, $\uparrow$*)
Frontal	δ (*, $\uparrow$*), θ (**, $\downarrow$**), γ (*, $\uparrow$*), β (*, $\uparrow$*)
Central	γ (**, $\uparrow$*), β (**, $\uparrow$**)
Partial	γ (**, $\uparrow$*), β (**, $\uparrow$*)
L-temporal	γ (*, $\uparrow$*), β (*, $\uparrow$*)
R-temporal	γ (*, $\uparrow$*), β (**, $\uparrow$*)
Occipital	γ (**, $\uparrow$**), β (**, $\uparrow$**)

"somewhat useful" (usefulness=2). Then we conduct exploratory analyses and usefulness estimation experiments on these results.

4 BRAIN SIGNALS ANALYSES

To address **RQ1**, we investigate the effects of search results' usefulness on the corresponding brain signals.

4.1 Statistical methods

We access the effects of search results' usefulness on the human brain by analyzing the spectral powers in different frequency bands and EEG channels, which are widely applied to measure brain activities [35]. For each EEG channel, we extract the spectral powers of each epoch between 0.5 and 50 Hz according to the Welch's method and average them over the frequency bands of delta (0.5-4Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma (30-50 Hz). In order to test the difference of spectral powers between brain signals in response to search results of different usefulness, we applied repeated measures ANOVA. The independent variable is the usefulness rating of search results. The dependent variable is each participants' mean spectral power in a frequency band and a EEG channel. After that, we apply post-hoc Bonferroni tests to conduct pair-wise comparisons between groups. We report ANOVA results in fine-grained and coarse-grained strategies regarding the division of brain regions. The fine-grained strategy treats each channel as a brain region. The coarse-grained strategy aggregate the adjacent channels' spectral powers in different brain regions (i.e., pre-frontal, frontal, central, parietal, l-temporal, r-temporal, and occipital) according to the 10-20 system [11].

Besides, to explore the mixed effects, such as the display type of the search result, the task order, and the word number in a search result, we conduct mixed effect analyses with a mixed linear model. We find that the effect of the search results' usefulness is significant when taking these confounding factors into account, which suggests that we can infer the search results' usefulness with brain signals robustly. The mixed effects analyses are elaborated in Section A.2.

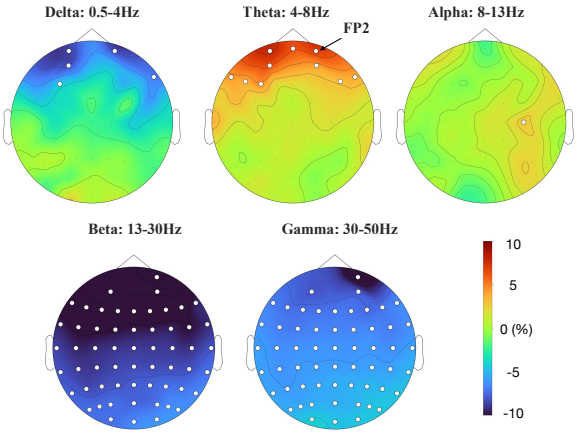


Figure 3: The relative differences of the spectral powers between groups of "not useful at all" and "very useful". The highlighted sensors indicate the differences of spectral powers in 1) the repeated measures ANOVA test regrading all groups and 2) the post hoc Bonferroni tests between groups of "not useful at all" and "very useful" are both significant at the $p < 0.05$ level.

4.2 Relationship between the EEG spectral power and result usefulness

Significance levels of spectra powers' differences among different usefulness ratings are reported in Table 2 and Figure 3. From Table 2, we observe that the usefulness ratings have significant effects on all the coarse-grained brain regions. Among all coarse-grained brain regions, the theta band power in pre-frontal achieves the most significant statistic in the ANOVA test (F[2,36]=12.67, p<1e-3) and the post-hoc test (M_{diff}=-5.96, p<1e-3). Besides, among all the EEG channels, the theta band power in FP2 (see in Figure 3) achieves the most significant statistic in the ANOVA test (F[2,36]=9.42, p<1e-3) and the post-hoc test (M_{diff}=-5.46, p<1e-3). Therefore, we suggest that brain signals are effective and robust factors in predicting usefulness judgment.

Additionally, we find that more significant channels have appeared in beta and gamma bands than others. The beta and gamma bands are related to stressed and alert levels [31]. We observe that the non-click search results with higher usefulness usually have lower spectral powers in these bands. Thus it suggests that the useless search result may cause negative emotion such as stress and anxious thinking. Besides, previous studies have revealed that the spectral powers in beta and gamma bands are more effective for emotion recognition [20]. In the scenario of search result examination, we speculate that whether the IN is realized and whether the user is satisfied might arouse patterns of advanced cognitive functions similar to certain positive emotions. Our findings indicate that we can detect positive emotions when the user is visiting useful results.

Significant findings also exist in delta and theta bands. Theta bands are related to cognitive and memory performance [19] while delta bands are traditionally considered to be associated with deep sleep [2]. We observe that the dominant band (i.e., band with

Table 3: The content, context, and brain signals features.

Information source	Features
Content	BM25 score, BERT score, result type
Context	result position, avg/max/total usefulness of previous results, avg/max similarity score with previous results
Brain signals	62 channels × (5 spectral domain features + 62 temporal domain features)

higher spectral power) is the theta band, which suggests the process related to working memory are more active in useful search results. Another observation is that the significant findings are distributed mainly at brain regions of frontal and pre-frontal. Previous functional magnetic resonance imaging (fMRI) studies on relevance perception [28] suggest that the brain activities at these regions (i.e., frontal and pre-frontal), are different when processing relevant and non-relevant documents. These findings indicate that relevance and usefulness, though differentiated by some researchers, might be two highly connected concepts sharing similar cerebral function areas.

*Answer to **RQ1**.* We analyze the effects of the usefulness of non-click results on the spectral powers of EEG signals. The above analyses provide converging and insightful evidence that there are detectable differences in brain activities while examining non-click search results of different usefulness ratings. It suggests that brain signals are effective in inferring usefulness judgment.

5 USEFULNESS ESTIMATION

To answer **RQ2**, we explore the effectiveness of brain signals in usefulness estimation. We conduct experiments to compare different models based on brain signals, content/context information, and their combinations. Furthermore, we analyze the effect of brain signals in different experimental settings.

Since the aim of this experiment is to demonstrate the effectiveness and robustness of EEG signals as feedback, we apply prevalent feature engineering methods and several state-of-the-art multichannel EEG classification models. The investigation on designing more sophisticated EEG classification models and methods to combine various information sources are left as future work.

5.1 Experimental setups

5.1.1 Features. This subsection elaborates selected features based on brain signals and content/context factors. For content and context features, we inherit the factors from [25, 25]'s study. Their study investigate the factors that affect usefulness judgments, e.g., BM25 score and result position, which is detailed in Section A.1 (e.g.,). For brain signals, existing works in multichannel EEG-based prediction extract features in the spectral domain and the temporal domain [9, 44]. In our practice, we extract differential entropy (DE) [13] as spectral domain features and down-sampling the raw EEG data of each channel to 25Hz as temporal domain features. DE is equivalent to the logarithm of band power as described in Section 4.1, which is considered to have a better performance than band power in EEG-based prediction [8].

5.1.2 Models. In general, EEG classification models can be divided into topology-invariant and topology-aware. Traditional classification models, such as support vector machines (SVM), k-Nearest Neighbors (KNN), and Gradient Boosting Decision Tree (DT), are belong to the group of topology-invariant. They do not consider the topological structure of EEG channels when extracting the information and adopt manually designed features, especially spectral features, to circumvent the issue of high dimensionality. In contrast, topology-aware classifiers, such as CNN [22], GNN [38], and attention-based model [15, 40], take the spatial relations of EEG channels into account and learn EEG representations by aggregating features from different channels.

To verify the effectiveness of brain signals, we exploit prevalent models from both groups. For the group of topology-invariant classifiers, we adopt DT, which is widely used in machine learning tasks since it can automatically choose and combine the EEG features. For topology-aware models, we exploit Graph Convolutional Neural Network (GCN) [38], Hierarchical Spatial Learning Transformer (HSLT) [40], and SST-EmotionNet (SST) [15]. GCN constructs heterogeneous graph to learn deep-level information of graph-structured EEG signals. HSLT and SST applies attention mechanisms to adaptively capture discriminative patterns in spectral and temporal information, which achieves state-of-art performance for EEG-based prediction tasks.

As for the modeling of content and context features, we compare the performance of DT, multilayer perceptron (MLP), and SVM in our dataset. Among them, DT achieves the best performance, which is consistent with previous work [26]. Due to the page limits, we only report the experimental results of DT.

Finally, we perform a grid search using a trade-off parameter λ (19 values from .05 to .95) to combine the estimation scores of models based on content/context information and brain signals. Since we aim to explore the effectiveness and robustness of brain signals, designing more sophisticated combination methods is left as future work.

5.1.3 Protocols and Evaluation. Given content/context features, brain signals features, or their combination, the task is to estimate the usefulness level of non-click search results. We simplify the task as a binary classification problem by only considering usefulness ratings of 1 and 4. The reasons are two-fold: (1) Ratings of 1 ("not useful at all") and 4 ("very useful") are boundary usefulness judgments, and thus they contain less noise than ratings of 2 and 3. (2) Ratings of 1 and 4 make up of 71.2% search results in total.

To verify the performance in different application scenarios, we perform two protocols in our experiments: *task-independent* and *user-independent*. The task-independent protocol partitions the tasks into ten folds then uses the rest folds for training when validating each fold. The user-independent protocol uses data of an individual participant for evaluation and trains with the remaining participants' data.

As for evaluation metrics, we follow the same principle as in [26] and also use Area Under Curve (AUC) for our task and report the standard deviation of AUC among different folds.

5.1.4 Parameter setups. For all models, the parameters are tuned according to the averaged AUC. For DT, the parameters include learning rate, estimator number, leaf nodes, and maximum tree depth, then the hyper parameters are selected from $\{10^{-4}, 10^{-3}, 10^{-2}\}$, $\{100, 200, 400\}$, $\{3, 9\}$, and $\{3, 9\}$, respectively. For GCN, HSLT, and SST, we inherit most of the hyper parameters from the original paper [15, 38, 40] and tune the parameters of learning rate and batch size, which are selected from $\{10^{-4}, 10^{-3}, 10^{-2}\}$ and $\{4, 8, 16\}$, respectively. Besides, to accelerate the training procedure, we train GCN, HSLT, and SST on an NVIDIA TITAN XP 12G GPU and adopt the early-stop strategy when the validation performance does not improve after five iterations.

5.1.5 Definitions. To avoid ambiguity, we use M^f to denote the model M (=DT, SST) using features f (= *cn*, *cx*, *bs*). *cn*, *cx*, and *bs* indicate content features, context features, and brain signal features, respectively. + denotes the combination of different models with trade-off parameters λ. For instance, $DT^{cn,cx} + SST^{bs}$ denotes the combination model of DT using features of content and context features and SST using brain signals.

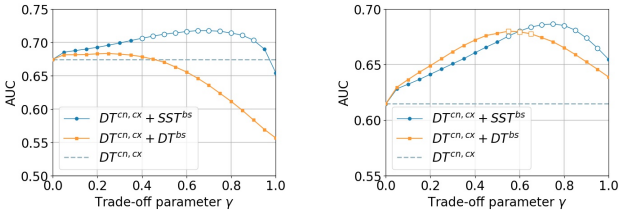
5.2 Results and Analysis

In this section, we report experimental results to answer **RQ2**. Primarily, we elaborate the overall performance of usefulness estimation with different information sources to demonstrate the effectiveness of brain signals. Additionally, we provide extensive analysis to study the effect of different experimental settings (i.e., trade-off parameter, task difficulty, and length of time interval).

5.2.1 Overall Performance. Table 4 shows the overall performance of the usefulness estimation of different models on the basis of various sources (i.e., content, context, brain signals, and their combination). For combination models, we mainly discuss $DT^{cn,cx} + DT^{bs}$ and $DT^{cn,cx} + SST^{bs}$ since DT^{bs} and SST^{bs} are representative topology-invariant and topology-aware model, respectively. And

Table 4: The performance of usefulness estimation with different information sources. M^f denotes model M using features f. *cn*, *cx*, and *bs* indicate content, context, and brain signals, respectively. + denotes grid search combination. $*/**$ indicate the difference of performance with $DT^{cn,cx} + SST^{bs}$ is significant with p-value < 0.05/0.01.

Model	task-independent		user-independent	
	AUC	STD	AUC	STD
DT^{cx}	0.585**	0.049	0.664**	0.047
DT^{cn}	0.593**	0.080	0.619**	0.040
$DT^{cn,cx}$	0.614*	0.067	0.672**	0.049
DT^{bs}	0.642	0.033	0.585**	0.047
GCN^{bs} [38]	0.644	0.030	0.591**	0.023
$HSLT^{bs}$ [40]	0.654	0.043	0.620**	0.030
SST^{bs} [15]	0.655	0.037	0.654**	0.043
$DT^{cn,cx} + DT^{bs}$	0.683	0.049	0.687**	0.049
$DT^{cn,cx} + SST^{bs}$	**0.687**	0.050	**0.718**	0.040



(a) User-independent. **(b) Task-independent.**

Figure 4: The performance of usefulness estimation with different trade-off parameter γ. When γ = 0, the performance coincides with $DT^{cn,cx}$. When γ = 1, the performance coincides with DT^{bs} or SST^{bs}. The hollow dot denotes the performance is significantly better than $DT^{cn,cx}$.

SST^{bs} outperforms other topology-aware models. From Table 4, we have the following observations:

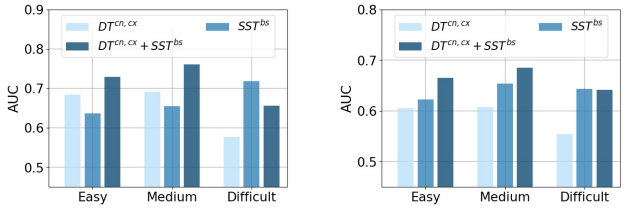
(1) For both protocols, models utilizing all features perform significantly better than models that ignore brain signals. The best performance is achieved by $DT^{cn,cx} + SST^{bs}$, in which we use SST for brain signals modeling and combine it with $DT^{cn,cx}$. This observation demonstrates that brain signals complement conventional features, including content and context features, and benefit usefulness estimation.

(2) For EEG models, the performance in user-independent protocol is worse than that in task-independent protocol, especially for DT. The reason is that the brain signals have individual differences [46]. In spite of this, our experimental results suggest that individual differences can be alleviated with the deep network of SST, which performs better than other models in user-independent protocol. The result highlights the effectiveness of the multi-stream attention mechanism in SST [15]. This finding is interesting, and we left the study of how to utilize brain signals to perform stability across protocols as future work.

(3) As for models excluding brain signals, they perform worse in the protocol of task-independent than user-independent. The reason is that some of the content and context features (e.g., BM25 score, total usefulness of previous results) are associated with the task, as discussed in Section A.1. Thus, the performance degrades for unseen tasks. However, models using brain signals do not perform worse in task-independent protocol than in user-independent protocol since they directly capture user's psychological feedback, which is not associated with the tasks.

5.2.2 In-depth Analysis.

Analysis of trade-off parameter. By using a trade-off parameter γ to combine the scores estimated by information sources of brain signals and content/context features, we aim to test (1) for which settings of γ the combination model performs better than $DT^{cn,cx}$ significantly and (2) whether the combination model is sensitive to the γ or not. In Figure 4, we show the performance of models using all features ($DT^{cn,cx} + DT^{bs}$ and $DT^{cn,cx} + SST^{bs}$) with different trade-off parameter γ. Recall that for γ = 0 and γ = 1, models degrades to the $DT^{cn,cx}$ and SST^{bs}/DT^{bs}, respectively. Since SST performs better for brain signals modeling, we mainly discuss the $DT^{cn,cx} + SST^{bs}$ and have two main observations.



(a) User-independent. (b) Task-independent.

Figure 5: The performance of usefulness estimation with various task difficulties. $DT^{cn,cx}$ performs worse in difficult tasks than that in easy and medium tasks ($p < 0.05$).

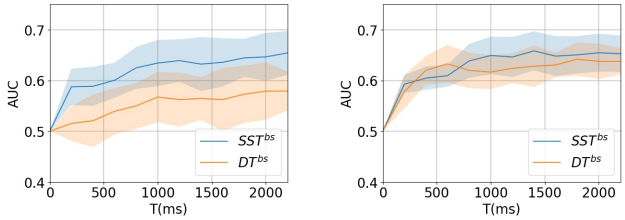
On the one hand, as γ increases, $DT^{cn,cx} + SST^{bs}$ monotonically increases to the best performance with an optimal value of γ and then gradually decreases. This finding demonstrates that incorporating conventional features and brain signals together is better than considering one facet only.

On the other hand, $DT^{cn,cx} + SST^{bs}$ is significantly better than $DT^{cn,cx}$ for $0.4 \leq \gamma \leq 0.85$ (user-independent) and $0.55 \leq \gamma \leq 0.9$ (task-independent). However, changing γ in $0.15 \leq \gamma \leq 0.85$ shows no significant differences (for both protocols). These suggest that the combination model is not sensitive to this parameter.

Analysis of task difficulty. The task difficulty collected in the user study is classified into three groups: easy (very easy and easy), medium (neither easy nor difficult), and difficult (difficult and very difficult). Then we calculate the performance of usefulness estimation across these groups, as shown in Figure 5.

The performance of $DT^{cn,cx}$ is worse in the difficult tasks than that in the easy tasks in both protocols. Especially in the protocol of user-independent, repeated measures ANOVA shows that there exists a significant difference among task difficulty levels ($F[20, 2] = 4.24, p < 0.05$). In contrast, the performance of models based on brain signals does not decrease along with the increase of task difficulty. This finding indicates that models using brain signals are effective and robust in difficult tasks.

Analysis of time intervals. Since brain activities are time-sensitive, we further explore the influences of the lengths of time intervals of brain signals on the model performance. Figure 6 shows the experimental results of SST^{bs} and DT^{bs}. We find no significant difference in terms of the model performance after 800ms in both protocols and both models. It is consistent with existing work that



(a) User-independent. (b) Task-independent.

Figure 6: The performance of usefulness estimation using brain signals with different time intervals of $[0, T]$.

suggests that our brain needs around 800ms to judge the relevance of a visually presented stimulus [1].

*Answer to **RQ2**.* According to the experimental results, we find that incorporating brain signals can improve the performance of usefulness estimation significantly (e.g., 7.3% in terms of AUC in the task-independent protocol). Besides, we verify the robustness of brain signals in different situations (i.e, unseen users and unseen tasks) and various experimental settings (i,e., trade-off parameter, task difficulty, and time intervals). The findings suggest that improving usefulness estimation with brain signals is beneficial.

6 SEARCH RESULT RE-RANKING

In this section, we devise two re-ranking methods, i.e., a PM and a GIM, to answer **RQ3**. The predicted usefulness in Section 5 is inherited in PM and GIM for search result re-ranking.

6.1 Problem Statement

We formulate the pth user's usefulness judgment within the tth search task on the ith search result as $I_p^{t,i} = \{d_p^{t,i}, u_p^{t,i}\}$, where $u_p^{t,i}$ denotes the usefulness score of search result $d_p^{t,i}$. Note that the search results in a certain search task are different for different users, i.e., $\langle d_{p_j}^{t,j_1}, d_{p_j}^{t,j_2}, ...\rangle \neq \langle d_{p_k}^{t,k_1}, d_{p_k}^{t,k_2}, ...\rangle$, since the result lists are shuffled and the users can break their search at any time (see Section 3.3). Then the re-ranking methods aim to rank results with higher usefulness judgment score to top positions given a user and a task.

6.2 Methods

6.2.1 Baselines. We adopt two baseline models for comparison: a probabilistic retrieval model BM25 [36] and a pre-trained model BERT (for text ranking [23]). These models only take the content information of queries and search results into account while leave out the search context.

6.2.2 PM. The personalized method re-ranks the search results with each person's predicted usefulness score. The re-ranked list Π_p^t can be formulated as:

$$\Pi_p^t = \S(\langle d_p^{t,i_1}, d_p^{t,i_2}, ...\rangle, \langle \hat{u}_p^{t,i_1}, \hat{u}_p^{t,i_2}, ...\rangle)$$

where $\hat{u}_p^{t,i}$ is the usefulness score predicted by the usefulness estimation models, § indicates the function to sort $\langle d_p^{t,i_1}, d_p^{t,i_2}, ...\rangle$ according to the predicted usefulness score $\langle \hat{u}_p^{t,i_1}, \hat{u}_p^{t,i_2}, ...\rangle$.

6.2.3 GIM. The generalized intent modeling method builds an intent estimation model to generate an intent representation with the wisdom of general users, which is shown in Algorithm 1. Given the user p_i and task t, we firstly generate the intent vector $\vec{I}$ by aggregating the search result vectors $\langle D_{p_j}^{t,j_1}, D_{p_j}^{t,j_2}, ...\rangle, j \neq i$ according to their predicted usefulness score $\langle \hat{u}_{p_j}^{t,j_1}, \hat{u}_{p_j}^{t,j_2}, ...\rangle, j \neq i$, where $D_{p_j}^{t,j}$ is the representation of $d_{p_j}^{t,j}$ using a pre-trained "bert-chinese-base" encoder [6]. The motivation for constructing the intent vector is that the usefulness score indicates the contribution of the search result to satisfy the intent. Then we generate the search result score $s_{p_i}^{t,i}$ for the user p_i and task t by calculating the cosine similarity

of intent vector $\vec{I}$ and the search result vector $D^{t,i}_{p_i}$. Finally, the re-ranked list $\Pi^t_{p_i}$ can be calculated by sorting the search results according to their scores $\langle s^{t,i_1}_{p_i}, s^{t,i_2}_{p_i}, ...\rangle$.

Algorithm 1: Generalized Intent modeling Method (GIM)

Input: Search results for the validating user $\langle d^{t,i_1}_{p_i}, d^{t,i_2}_{p_i}, ...\rangle$;
Search results for the other users $\langle d^{t,j_1}_{p_j}, d^{t,j_2}_{p_j}, ...\rangle$;
The estimated usefulness scores $\langle \hat{u}^{t,j_1}_{p_j}, \hat{u}^{t,j_2}_{p_j}, ...\rangle$, $j \neq i$.

Data:
Score list $\langle s^{t,i_1}_{p_i}, s^{t,i_2}_{p_i}, ...\rangle$; Intent vector $\vec{I}$;
Search result vector $\vec{D^{t,i}_p}$;

Init;
$s^{t,i}_{p_i} = 0$; $\vec{I} = \vec{0}$; $\vec{D^{t,i}_p} = BERT(d^{t,i}_p)$;
$\overline{\hat{u}^t_{p_i}} = average\{\hat{u}^{t,j_k}_{p_j} \in \langle \hat{u}^{t,j_1}_{p_j}, \hat{u}^{t,j_2}_{p_j}, ...\rangle, j \neq i\}$;
for *all* $j \neq i$ **do**
  for *all* $d^{t,j_k}_{p_j} \in \langle d^{t,j_1}_{p_j}, d^{t,j_2}_{p_j}, ...\rangle$ **do**
    $\vec{I} = \vec{I} + D^{t,j_k}_{p_j} \cdot (\hat{u}^{t,j_k}_{p_j} - \overline{\hat{u}^t_{p_i}})$
  end
end
for *all* $s^{t,i_k}_{p_i} \in \langle s^{t,i_1}_{p_i}, s^{t,i_2}_{p_i}, ...\rangle$ **do**
  $s^{t,i_k}_{p_i} = cosine\_similarity(\vec{I}, D^{t,i_k}_{p_i})$
end
$\Pi^t_{p_i} = \S(\langle d^{t,i_1}_{p_i}, d^{t,i_2}_{p_i}, ...\rangle, \langle s^{t,i_1}_{p_i}, s^{t,i_2}_{p_i}, ...\rangle)$;
return $\Pi^t_{p_i}$;

6.3 Experimental settings

In our experiment, we inherit the usefulness score of $\text{DT}^{cn,cx}$, SST^{bs}, and $\text{DT}^{cn,cx} + \text{SST}^{bs}$ since they perform better than other models. The experiments using the usefulness score in the task-independent and user-independent protocols show consistent findings in the comparisons of the models. Thus, we only present observations in the task-independent protocol. To avoid ambiguity, we use M^f to denote the model M (=BM25, BERT, PM, GIM) using features f (= cn, cx, bs). For example, $BM25^{cn}$ indicates the BM25 model utilizing the content features and $GIM^{cn,cx,bs}$ indicates the GIM models inheriting the usefulness score of $\text{DT}^{cn,cx} + \text{SST}^{bs}$.

To compare the performance of different models and features, we utilize two popular evaluation metrics: Normalized Discounted Cumulative Gain (NDCG) [14] and Mean Reciprocal Rank (MRR) [33]. Since the average amount of documents in a search task is 3.41, we calculate NDCG at different cutoff positions of $\{1, 3, 5\}$, i.e., $NDCG@\{1, 3, 5\}$. And we report MRR of the full ranked list.

6.4 Results and discussions

Table 5 shows the ranking performance of our re-ranking methods using different features. From Table 5, we have the following observations:

(1) $BERT^{cn}$ performs better than $BM25^{cn}$ in terms of most evaluation metrics, i.e., $NDCG@3$, $NDCG@5$, and MRR. However, their performance is significantly worse than all PM and GIM models in all evaluation metrics. The reason is that the result usefulness can't be simply judged with semantic score only [3, 37].

(2) For PM and GIM, models using additional information of brain signals (i.e., $PM^{cn,cx,bs}$ and $GIM^{cn,cx,bs}$) perform significantly better than models using content and context information only (i.e., $PM^{cn,cx}$ and $GIM^{cn,cx}$), respectively. This result demonstrates that utilizing brain signals can improve re-ranking performance.

(3) $GIM^{cn,cx,bs}$ performs significantly better than $PM^{cn,cx,bs}$, which indicates our intent modeling method is more effective. Additionally, the GIM re-ranks the result list by modeling the search intent in the corresponding search task. Therefore, it can be adopted to unseen users and unseen search results. In contrast, the PM generates a re-ranked list for search results with predicted usefulness scores, and thus it is unpractical for unseen search results. Besides, the PM underperforms GIM since PM suffers from the problem of unstableness in modeling the search behavior and the brain activities of only an individual user. Nevertheless, it is worth mentioning that the PM has its advantages since it takes personalized information into account. It is interesting to conduct future work to design a more sophisticated PM and analyze its effectiveness in personalized IR scenarios, such as personalized search or recommendation.

*Answer to **RQ3**.* According to the experimental results, we can observe that PM and GIM perform significantly better than the baselines. Besides, brain signals as additional information sources can significantly improve the performance of both PM and GIM.

Table 5: The performance of search result re-ranking with different information sources. M^f denotes model M using features f. *cn, cx,* and *bs* indicate content, context, and brain signals, respectively. $*$ indicate the difference of performance with $GIM^{cn,cx,bs}$ is significant with p-value < 0.01.

Model	NDCG@1	NDCG@3	NDCG@5	MRR
$BM25^{cn}$	0.407*	0.672*	0.725*	0.621*
$BERT^{cn}$	0.399*	0.691*	0.737*	0.655*
$PM^{cn,cx}$	0.446*	0.714*	0.751*	0.677*
PM^{bs}	0.457*	0.725*	0.764*	0.691
$PM^{cn,cx,bs}$	0.522*	0.752*	0.787*	0.726*
$GIM^{cn,cx}$	0.490*	0.739*	0.775*	0.709*
GIM^{bs}	0.571	0.776	0.811	0.754
$GIM^{cn,cx,bs}$	**0.591**	**0.787**	**0.814**	**0.764**

7 CONCLUSIONS AND DISCUSSIONS

Understanding the non-click results is increasingly significant with the growing percentage of SERP snippets satisfying the user's information need directly. In this paper, we design a user study and analyze the relationship between brain signals and the usefulness of non-click search results. We find detectable differences in various spectral bands and brain regions. Additionally, our neurology analysis indicates that usefulness judgments are associated with several cognitive functions related to positive emotions, working memory, and relevance perception.

Inspired by the findings above, we conduct extensive experiments on usefulness estimation and search result re-ranking for

non-click results based on brain signals and conventional factors, i.e., content and context factors. Insightful findings include: (1) brain signals are effective for usefulness estimation and are more robust than conventional features in different protocols and experimental settings; (2) the performances of models only using conventional features degrades in difficult tasks while models based on brain signals do not; (3) the search result re-ranking performance is significantly improved with the usefulness estimated models using brain signals.

With the development of wearable devices, researchers have already built an available system of using BCI to replace keyboard and mouse in search scenarios [4, 24]. On top of that, our research shows additional promising benefits with the practical application of BCI in search engines. Besides helping applications in which hand-based interactions are infeasible, we suggest that the benefits of BCI for the search system are two-fold: (1) BCI can detect user satisfaction with real-time brain signals. To make the first move in BCI-enhanced IR, we study the non-click results, which are special circumstances where conventional user interactions (i.e., click, dwell time in the landing page) are inaccessible. We believe this paradigm can also improve performance in other scenarios lacking feedback signals. (2) BCI has advantages over traditional search feedback. Traditional web search heavily relies on implicit feedback (e.g., click, dwell time) and explicit feedback (i.e., human annotation) to improve performance. Since BCI directly captures brain activities, we suggest using brain signals as special "explicit feedback", which requires no extra efforts for annotation and represent real search experience. With the BCI devices becoming portable in the near future, it is promising to utilize brain signals for search evaluation and performance improvement.

Several limitations guide exciting directions for future work: (1) In this paper, we perform lab-based settings in our user study. Analyzing brain signals in real-life search scenarios with portable EEG devices and taking temporal and demographic aspects into account is an interesting future work. (2) We demonstrate that brain signals are valuable for the usefulness estimation and can be collected almost in real-time. Hence, designing scene-adaptive methods to model brain signals and combine various information sources for real-time proactive IR systems is a promising direction.

8 ACKNOWLEDGEMENT

This work is supported by the Natural Science Foundation of China (Grant No. 61732008), Beijing Academy of Artificial Intelligence (BAAI), and Tsinghua University Guoqiang Research Institute.

REFERENCES

[1] Marco Allegretti, Yashar Moshfeghi, Maria Hadjigeorgieva, Frank E Pollick, Joemon M Jose, and Gabriella Pasi. 2015. When relevance judgement is happening? An EEG-based study. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval*. 719–722.

[2] F Amzica and M Steriade. 1998. Electrophysiological correlates of sleep delta waves. *Electroencephalography and clinical neurophysiology* 107, 2 (1998), 69–83.

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

[4] Xuesong Chen, Ziyi Ye, Xiaohui Xie, Yiqun Liu, Weihang Su, Shuqi Zhu, Min Zhang, and Shaoping Ma. 2021. Web Search via an Efficient and Effective Brain-Machine Interface. arXiv:arXiv:2110.07225

[5] 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.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Abdigani Diriye, Ryen White, Georg Buscher, and Susan Dumais. 2012. Leaving so soon? Understanding and predicting web search abandonment rationales. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. 1025–1034.

[8] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. 2013. Differential entropy feature for EEG-based emotion classification. In *6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 81–84.

[9] Manuel JA 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.

[10] 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.

[11] Richard W Homan, John Herman, and Phillip Purdy. 1987. Cerebral location of international 10–20 system electrode placement. *Electroencephalography and clinical neurophysiology* 66, 4 (1987), 376–382.

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

[13] A Hyvärinen. 1998. New approximations of differential entropy for independent component analysis and projection pursuit. *Advances in neural information processing systems* 10 (1998), 273.

[14] Kalervo Järvelin and Jaana Kekäläinen. 2017. IR evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 243–250.

[15] Ziyu Jia, Youfang Lin, Xiyang Cai, Haobin Chen, Haijun Gou, and Jing Wang. 2020. Sst-emotionnet: Spatial-spectral-temporal based attention 3d dense network for eeg emotion recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2909–2917.

[16] Thorsten Joachims et al. 2003. Evaluating Retrieval Performance Using Clickthrough Data.

[17] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR Forum*, Vol. 51. Acm New York, NY, USA, 4–11.

[18] Hyun Hee Kim and Yong Ho Kim. 2019. ERP/MMR algorithm for classifying topic-relevant and topic-irrelevant visual shots of documentary videos. *Journal of the Association for Information Science and Technology* 70, 9 (2019), 931–941.

[19] Wolfgang Klimesch. 1999. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain research reviews* 29, 2-3 (1999), 169–195.

[20] Sander Koelstra, Christian 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.

[21] 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.

[22] Jinpeng Li, Zhaoxiang Zhang, and Huiguang He. 2018. Hierarchical convolutional neural networks for EEG-based emotion recognition. *Cognitive Computation* 10, 2 (2018), 368–380.

[23] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. Pretrained transformers for text ranking: Bert and beyond. *Synthesis Lectures on Human Language Technologies* 14, 4 (2021), 1–325.

[24] Yiqun Liu, Jiaxin Mao, Xiaohui Xie, Min Zhang, and Shaoping Ma. 2021. Challenges in designing a brain-machine search interface. In *ACM SIGIR Forum*, Vol. 54. ACM New York, NY, USA, 1–13.

[25] Jiaxin Mao, Yiqun Liu, Noriko Kando, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. Investigating result usefulness in mobile search. In *European Conference on Information Retrieval*. Springer, 223–236.

[26] Jiaxin Mao, Yiqun Liu, Huanbo Luan, Min Zhang, Shaoping Ma, Hengliang Luo, and Yuntao Zhang. 2017. Understanding and predicting usefulness judgment in web search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1169–1172.

[27] Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. 2016. When does relevance mean usefulness and user satisfaction in web search?. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 463–472.

[28] Yashar Moshfeghi, Luisa R Pinto, Frank E Pollick, and Joemon M Jose. 2013. Understanding relevance: An fMRI study. In *European conference on information retrieval*. Springer, 14–25.

[29] Yashar Moshfeghi, Peter Triantafillou, and Frank Pollick. 2019. Towards predicting a realisation of an information need based on brain signals. In *The World Wide Web Conference*. 1300–1309.

[30] 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.

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

[32] 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.

[33] Dragomir R Radev, Hong Qi, Harris Wu, and Weiguo Fan. 2002. Evaluating Web-based Question Answering Systems.. In *LREC*. Citeseer.

[34] Filip Radlinski and Thorsten Joachims. 2005. Query chains: learning to rank from implicit feedback. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. 239–248.

[35] Brian J Roach and Daniel H Mathalon. 2008. Event-related EEG time-frequency analysis: an overview of measures and an analysis of early gamma band phase locking in schizophrenia. *Schizophrenia bulletin* 34, 5 (2008), 907–926.

[36] Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94*. Springer, 232–241.

[37] Ian Ruthven. 2021. Resonance and the experience of relevance. *Journal of the Association for Information Science and Technology* 72, 5 (2021), 554–569.

[38] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. 2018. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing* 11, 3 (2018), 532–541.

[39] Ellen M Voorhees. 2001. The philosophy of information retrieval evaluation. In *Workshop of the cross-language evaluation forum for european languages*. Springer, 355–370.

[40] Zhe Wang, Yongxiong Wang, Chuanfei Hu, Zhong Yin, and Yu Song. 2022. Transformers for EEG-based emotion recognition: A hierarchical spatial information learning model. *IEEE Sensors Journal* (2022).

[41] Zhihua Wang, Yang Yu, Ming Xu, Yadong Liu, Erwei Yin, and Zongtan Zhou. 2019. Towards a hybrid BCI gaming paradigm based on motor imagery and SSVEP. *International Journal of Human–Computer Interaction* 35, 3 (2019), 197–205.

[42] Kyle Williams, Julia Kiseleva, Aidan C Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. 2016. Detecting good abandonment in mobile search. In *Proceedings of the 25th International Conference on World Wide Web*. 495–505.

[43] 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.

[44] Ziyi Ye, Xiaohui Xie, Yiqun Liu, Zhihong Wang, Xuesong Chen, Min Zhang, and Shaoping Ma. 2021. Understanding Human Reading Comprehension with brain signals. arXiv:arXiv:2108.01360

[45] 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.

[46] Wei-Long Zheng and Bao-Liang Lu. 2016. Personalizing EEG-based affective models with transfer learning. In *Proceedings of the twenty-fifth international joint conference on artificial intelligence*. 2732–2738.

A SUPPLEMENTARY MATERIAL

A.1 Content and context factors

Among the literature of usefulness judgment, Mao et al. [26] firstly investigate the factors that affect usefulness judgments in desktop search scenarios and extend their study into mobile devices [25]. They conduct usefulness estimation task using two types of factors related to: (1) *content* information (i.e., factors of current search result) and (2) *context* information (i.e., factors of interaction history). The factors in our experiment are mostly inherited from their study, excluding those related to interactions with the landing page, such as dwell time, scrolling, etc., which are not available for non-click results. In addition, we supplement the content factors with *result type*, which is intuitively related to the usefulness of non-click results. The factor *result type* consists of 19 categories according to their presentation styles, such as "Question Answering" and "Tutorial" [45]. Consequently, the *content* factors consist of *BM25 score* [36], *BERT score* [23], and *result type*. And we also utilize the *context* factors of *result position*, *average/max/total usefulness ratings with previous search results*, and *average/max similarity score with previous search results* from Mao et al. [26]'s study. In summary, the *content* and *context* factors are presented in Table 3.

For content factors, BM25 score and BERT score have positive effects on usefulness ratings. However, the effect (Pearson's r=0.086, p < 1e-3 for BERT score) is lower than that reported in Mao et al. [26]'s study (Pearson's r=0.18). The reasons are two-fold: 1) Text information provided on the SERP is less than that on the entire document. 2) The real-world top ten search results need more than text information to judge their usefulness [3]. Besides, result type of Question Answering (Pearson's r=0.127, p < 1e-3) plays a positives role since it can satisfy the user with its snippet directly. These findings indicate that utilizing text information for usefulness estimation is insufficient and demonstrate the benefits of enhanced search results, e.g., Question Answering type.

For context factors, the result position has negative correlation (Pearson's r=-0.095,p<1e-3) with usefulness. This finding suggests that the usefulness measures the increment of information when visiting result lists, and thus the usefulness of search results will diminish. Nevertheless, the total usefulness of previous results has no effect (Pearson's r=-0.017), which is different from previous research [26]. Mao et al. [26] observe that the total usefulness of previous results has a negative effect due to the redundancy with previous documents. But in our study, participants are allowed to break the search process once they are satisfied, and thus the redundancy problem is alleviated. However, most of the correlations of context factors are weak since some valuable interaction information such as click and dwell time do not exist in zero-click scenarios. This finding implies the demand for additional factors in zero-click search process.

A.2 Mixed effects analyses

In this section, we discuss some confounding factors that may affect the independence of our observations in Section 4. The confounding factors include: individual difference (I), the display type of the search result (D), the task order (O_t), the search result order (O_s), the word number in the search results (W). A linear mixed model is used for modeling the dependence of brain activities (B) measured by EEG spectral powers and the search result usefulness (U), which can be specified as:

$$B = (\beta_u + i_u)U + \beta_w W + \beta_t O_t + \beta_s O_s + \sum_{j=1,2,..,C-1} \beta_{d,j} D_j + I + \beta_0 + e$$

where e is the general residual error, β_0 is the general intercept, $\beta_u, \beta_w, \beta_t, \beta_s, \beta_{d,j}$ are coefficients corresponding to different effects. I is the individual difference effect and i_u is a random by-participant coefficient in respect to the search result usefulness. The word number in the search results (W) is a continuous variable. The task order (O_t) and the search result order (O_s) are continuous variables indicating the task rank for the participants and the search result rank within a task, respectively. Note that we randomize them in the data collection procedures. The display type of the search result (D) is a category variable, and C is the category number, which is detailed in Section A.1. Note that the category variable in the mixed linear model has only $C-1$ degrees of freedom. Then we model the dependence of brain activity (B) (estimated by spectral power, detailed in Section 4) and the search result usefulness (U) (ranging from 1 to 4).

Firstly, with the mixed linear model, we find significant correlations ($p < 0.05$) between the search result usefulness and the spectral power in 168 different channel-band pairs (e.g, the FP2 channel in the theta band). On the other hand, the prior analyses (see in Figure 3), which don't take the confounding factors into account, show 141 significant channel-band pairs, and 107 of them are identical to the analyses with the mixed linear model. This suggests that the dependence of brain activity and the search result usefulness are robust and are less affected by the above confounding factors. Then we discuss the effect of confounding factors on the FP2 channel in the theta band (the most significant finding in prior analysis in Section 4) in detail.

Table 6 presents the statistical results of the mixed linear model when measuring the brain activities with the spectral power of the FP2 channel in the theta band. From Table 6, we have the following observations: (1) The effect of usefulness on brain activities is the most significant. It verifies that when doing usefulness judgments, the brain activities are different regarding the usefulness ratings. (2) The effect of task order is not significant since we randomize the task order for each participant. But the search result order has a significant effect on brain activities. The search result order is indeed related to the usefulness judgments since the top search

Table 6: The statistical results of the mixed linear model. Coef. and z indicate the coefficient variable and the statistic corresponding to the effect, respectively.

Effects	Coef.	Std	z	p>z
Word number	0.000	0.000	1.339	0.180
Search result order	-0.012	0.004	-2.720	0.007
Task order	-0.001	0.000	-1.867	0.062
Display type (QA)	0.006	0.060	0.099	0.921
Display type (mixed)	0.046	0.022	2.096	0.036
Display type (organic)	0.036	0.028	1.319	0.187
Usefulness	0.029	0.006	**5.079**	**0.000**

results usually provide more useful information in the search process. It is not clear whether this effect of search result order can be separated from the search results' usefulness. (3)Some display types have significant effects, and three typical display types are selected to present on the table, i.e., QA, mixed, and organic. We find that the display type of mixed have a significant positive effect and its coefficient variable is higher than that of the organic type. We speculate the reasons are two-fold. On the one hand, mixed type contains both image and text, and thus it contributes to usefulness of the search result and further affects the brain activities. On the other hand, the content variability of different search result types may have an effect that can be separated from the search results' usefulness.

In general, although there exist several confounding factors in our study, we suggest the effect of usefulness on brain activities significantly exists. Therefore, we suggest that brain activities can be utilized to infer users' usefulness judgments.