Received: 3 March 2024 | Revised: 8 May 2024 | Accepted: 1 June 2024

DOI: 10.1002/asi.24936

RESEARCH ARTICLE

JASIST WILEY

Comparing point-wise and pair-wise relevance judgment with brain signals

Shuqi Zhu[1] | Xiaohui Xie[1] | Ziyi Ye[1] | Qingyao Ai[1,2] | Yiqun Liu[1,2]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China

[2]Zhongguancun Laboratory, Beijing, China

Correspondence
Qingyao Ai, Department of Computer Science and Technology, Zhongguancun Laboratory, Tsinghua University, Beijing, China.
Email: aiqy@tsinghua.edu.cn

Abstract

How to collect relevance judgment has long been an important problem in Information Retrieval (IR). A popular method is to collect relevance judgment in a point-wise manner, in which assessors examine and give an absolute relevance score for each item independently of the others. As an alternative, pair-wise relevance judgment, also named preference judgment, allows an assessor to compare two items side-by-side and express their preference for one over the other. Previous work has explored the differences between these two paradigms of relevance judgments from many different aspects. Most of these works are conducted through explicit/implicit feedback. However, few works investigate the underlying neurological mechanisms of the two paradigms. In this paper, we conduct a lab study to investigate and compare point-wise and pair-wise relevance judgment in image search scenarios. We study the neurological mechanisms of the two paradigms through an event-related potential (ERP) analysis of the users' brain signals while viewing images during a search process. We have obtained several observations, such as search engine users tend to pay more attention to preferred items in the point-wise paradigm but unpreferred items in the pair-wise paradigm. Furthermore, we test the adoption of brain signals as implicit feedback for predicting pair-wise relevance judgment, highlighting the feasibility of leveraging brain signals to understand users' relevance judgments.

1 | INTRODUCTION

Understanding the process of relevance judgment is an important problem in the field of Information Retrieval (IR). For a long time, offline evaluation of IR systems, following the Cranfield framework, has heavily relied on **point-wise relevance judgment** (Jain & Varma, 2011; Wu et al., 2020; Xie et al., 2018). Point-wise relevance judgment, also known as **graded relevance judgment**, is a method where assessors independently evaluate the relevance of an item using a graded scale (Cleverdon, 1967). However, point-wise relevance judgment has several drawbacks. For example, to the best of our knowledge, there is no universal grading scheme (i.e., how many levels to use and what those levels mean) in point-wise relevance judgment (Xie et al., 2020). Different numerical scales will significantly affect evaluation performance in various scenarios (Chu et al., 2021), as they determine the granularity of judgment and the interpretation of each level, which hurts the reliability of point-wise relevance judgment in practice.

As an alternative, **pair-wise relevance judgment**, also known as **preference judgment**, has recently drawn considerable attention in the research community. This paradigm allows assessors to compare two items simultaneously and express a preference for one over the

other (Carterette et al., 2008). Meanwhile, comparing two items side-by-side helps assessors make a faster and more accurate judgment since items in pairs can be seen as context to each other (Carterette et al., 2008; Clarke et al., 2021). Compared to assigning a numerical grade to items one by one, it is easier to recognize fine distinctions and express them in relative terms by collecting preferences directly (Yan et al., 2022).

Numerous prior works have investigated the differences and relationships between the two paradigms. Researches indicate that, despite the exponential increase in the number of item pairs to be judged with the total number of items, the pair-wise paradigm fosters greater consensus among assessors and streamlines the evaluation process by focusing on direct comparisons between pairs of items, ultimately resulting in higher inter-assessor agreement and reduced time consumption compared to the point-wise paradigm (Bah et al., 2015; Radinsky & Ailon, 2011). Also, when distinguishing the meaningful differences between two or more highly relevant items at a fine-grained level, preference judgment is usually more accurate (Yan et al., 2022). However, to the best of our knowledge, existing studies on this topic only compare the consistency and efficiency of the two paradigms through explicit feedback and behavior analysis. The actual neurological mechanisms behind the scenes are still unknown. The point-wise paradigm's lower consistency than the pair-wise paradigm is always attributed to the lack of a universally defined grading scheme. Understanding the reasons behind this inconsistency can aid in the design of improved grading schemes. Additionally, investigating why users find it easier and more efficient to make relevance judgments in the pair-wise paradigm than in the point-wise paradigm is of interest. With efforts being made to integrate pair-wise and point-wise relevance judgments for improved annotation efficiency (Chu et al., 2021; Yan et al., 2022), a neurological understanding of these paradigms can inform the design of more user-friendly annotation tasks.

In this paper, we aim to compare point-wise and pair-wise relevance judgment from a neuroscience perspective, explore the differences and similarities between two paradigms, and gain insights for IR tasks, thus raising the following research questions:

RQ1. What are the differences and similarities in the patterns of brain signals between pair-wise and point-wise relevance judgment?[1]

RQ2. What are the neurological mechanisms under these differences and similarities and how can these differences guide modern search techniques?

RQ3. To what extent can users' brain signals be used to predict their relevance judgment?

To investigate the above research questions, we use image search as an example application scenario. During the inspection of Search Engine Result Page (SERP) in web search, assessors typically examine multiple results on a single page and compare their usefulness to determine subsequent interactions, especially in image search scenarios where grid-based layouts present a larger number of results compared to sequential lists (Xie, Mao, Liu, de Rijke, Shao, et al., 2019; Xie et al., 2020). This is the reason we choose the image search scenario for our study. We conduct a lab-based user study in which human participants are required to accomplish a set of image relevance annotation tasks in the two paradigms separately (with a 7-day interval between them) and collect their brain signals in the tasks. Recent developments in neuroimaging technology (e.g., electroencephalogram [EEG] and functional magnetic resonance imaging [fMRI]) allow researchers to study brain activity patterns in IR scenarios in a more interactive way, leading to significant progress in understanding cognitive patterns that are difficult to reveal through traditional methods (Allegretti et al., 2015; Liu et al., 2021; Moshfeghi et al., 2016; Pinkosova et al., 2020). Through the use of EEG devices and event related potential (ERP) analysis (Luck et al., 2000), we find that significant differences in terms of brain activations and activities exist between point-wise and pair-wise relevance judgment. This implies that the process of the two paradigms involves distinct neurological functions. We also notice that the pair-wise paradigm leads to less cognitive load for assessors, and people tend to allocate more attention to items that align with their needs in the point-wise paradigm but pay more attention to comparatively inferior items in the pair-wise paradigm. Based on these findings, we provide suggestions for future IR tasks related to relevance judgment. Furthermore, we conduct several prediction experiments using frequency domain features of EEG signals. We verify the feasibility of employing EEG signals for predicting pair-wise relevance judgment and show that EEG signals can be utilized as implicit user feedback in IR.

2 | RELATED WORK

2.1 | Relevance judgment

Relevance judgment plays a vital role in IR system evaluation and ranking model optimization. The Cranfield-like framework (Cleverdon, 1967), in which assessors give a

graded relevance to each result returned by the retrieval system, has been widely used for search evaluation. Based on this framework, existing works have designed various grading schemes by assigning annotated graded relevance for each result. For example, Sang et al. (2011) employ binary judgment while Yang et al. (2015) and Clarke et al. (2004) use 3-level grading. More fine-grained schemes include 4-level grading by Luo et al. (2017) and Xie, Mao, Liu, de Rijke, Ai, et al. (2019), 6-level grading by Collins-Thompson et al. (2015), and even 100-level grading (S100) by Shao et al. (2019) and Roitero et al. (2018). They indicate that S100 is more consistent with the satisfaction of users than coarse-grained grading schemes. Traditional evaluation metrics based on point-wise relevance judgment, such as NDCG and RBP, have been widely used in IR.

In addition to the Cranfield-like approach and corresponding point-wise paradigms, several researchers have explored pair-wise relevance judgment and developing evaluation measurements. For instance, Carterette et al. (2008) investigate to evaluate search engines using preference judgment and design an interface for conducting such judgment. The study by Hui and Berberich (2017) finds that incorporating weak preference judgment, which includes a "tie" option, can help reduce evaluation costs. Furthermore, Carterette et al. (2008) propose two evaluation metrics, namely Ppref and Wpref, based on pair-wise relevance judgment. In the context of image search scenarios, Xie et al. (2020) introduce a novel preference-based evaluation metric called preference-winning-penalty (PWP).

Previous works have compared the consistency and efficiency of the two paradigms and have gained several insights. Research suggests that while the number of item pairs to be judged increases exponentially with the total number of items, the pair-wise paradigm achieves higher inter-assessor agreement and consumes less time compared to the point-wise paradigm (Bah et al., 2015; Radinsky & Ailon, 2011). Preference judgment has been recognized as a helpful approach to identifying meaningful differences between highly relevant items (Yan et al., 2022). Additionally, a comparison conducted by Yang et al. (2018) reveals that preference judgment is more reliable than other paradigms. Chu et al. (2021) propose a combined evaluation metric named pairwise discriminative power (PDP) to evaluate the quality of relevance judgment collections with both pair-wise signals and point-wise signals. A novel combined metric proposed by Arabzadeh et al. (2023) is applicable for instant search rather than offline search.

In addition, when inspecting SERP generated by a web search engine, assessors are expected to look through more than one result on a single page and may compare their usefulness to decide the next interaction. This phenomenon happens more frequently in image search scenarios as more results are presented on the SERP using a grid-based style other than a sequential list (Xie et al., 2020; Xie, Mao, Liu, de Rijke, Shao, et al., 2019). Image retrieval has consistently been a focal point of research and holds significant importance in real-world applications (Koh et al., 2024; Zhang et al., 2022). As the attractiveness and quality of images are becoming a more and more major factor for web image search engines to satisfy users' search intentions, assessors with different aesthetic standards probably give less consistent ratings (Geng et al., 2011). The limited demand for fresh results in web image search also enhances the reusability of pair-wise relevance judgment, which would have needed more amount of evaluation than the point-wise paradigm (Lefortier et al., 2014; Xie et al., 2020). For the reasons stated above, we choose image search scenarios as examples to conduct our study.

2.2 | BMI for search

In most of the previous studies on the comparison of the two paradigms, explicit/implicit feedback is used. Explicit feedback involves users explicitly stating their relevance judgments, which will increase the cognitive burden and user effort (Moshfeghi & Jose, 2013; White et al., 2002). Implicit feedback captures users' natural actions or physiological responses during system interactions, including measures like eye-tracking (Gwizdka et al., 2017) and search log analysis (Wu et al., 2019). However, implicit feedback often suffers from low signal-to-noise ratio (SNR) issues (Allegretti et al., 2015). Brain signals provide a direct reflection of psychological activities, making them a superior approach for perceiving human relevance judgments with greater immediacy and accuracy compared to other methods.

With the rapid developments of neuroimaging technology (e.g., EEG and fMRI), recent works have begun to study brain activity patterns in IR scenarios. For instance, Moshfeghi et al. (2016) employ fMRI to detect the emergence of Information Need (IN), while Allegretti et al. (2015) use EEG to explore the relevance judgment process. Furthermore, studies by Ye et al. (2022) utilize the ERP method to investigate the process of reading comprehension. In the context of point-wise relevance judgment, Pinkosova et al. (2020, 2022) conduct a user study and observe significant differences in ERP of high-relevance, low-relevance, and no-relevance answers. Pinkosova et al. (2023) suggest that self-perceived knowledge

(SPK) play an important role in relevance assessment, and Michalkova et al. (2024) explore the neurological mechanisms behind feeling-of-knowing. These researches on fundamental concepts make significant progress in the process of understanding cognitive patterns in IR scenarios, which is hard to reveal through previous technologies. However, the underlying differences in psychological activities related to relevance judgments of point-wise and pair-wise paradigms remain unknown.

Brain signals can also be utilized as a complement or substitute for traditional implicit feedback signals. Eugster et al. (2014) demonstrate the feasibility of detecting term relevance using brain signals, enabling the collection of relevance judgments without any additional user interactions. Moreover, Davis III et al. (2020) utilize the brain responses of a collective of human participants to improve the prediction of users' relevance judgments. Liu et al. (2021) propose to integrate BMI into the search system to understand users' information needs and collect direct satisfaction feedback.

3 | USER STUDY

In our user study, participants engage in simulated image search scenarios in point-wise and pair-wise paradigms sequentially, with a 7-day interval between them. The whole experimental process is carried out in the laboratory environment. This section describes the entire process of user data collection. The open-source of our experiment platform and dataset is available in the github.[2]

3.1 | Participants

We recruit a total of 20 participants through social media, including 15 males and 5 females. Participants are all college students aged between 19 and 25 years old and with a mean age of 21.7 and a standard deviation (SD) of 1.25 years. Their majors cover computer science, engineering, chemistry, law, and so on. Their education levels cover undergraduate and postgraduate. All participants are right-handed and claim to be proficient in using the Internet and search engines in their daily lives. It takes about 2 h to complete the pair-wise paradigm and 1.5 h to complete the point-wise paradigm, both including 30 min of equipment preparation and task guidance. Before the experiment starts, the participants are told to be paid US$11.8 per hour if they complete the experiment to ensure the quality of the collected user study data.

3.2 | Preparation

3.2.1 | Apparatus

The stimuli are presented on a desktop computer that has a 27-inch monitor with a resolution of 2560 × 1440 pixels and a refresh rate of 60 Hz. Participants are required to use the keyboard to interact with the platform. EEG signals are captured and amplified using a Scan NuAmps Express system (Compumedics Ltd., VIC, Australia) and a 64-channel Quik-Cap (Compumedical NeuroScan). A laptop computer functions as a server to record EEG signals and triggers using Curry8 software. Throughout the experiment, electrode-scalp impedance is maintained under 50 kΩ, and the sampling rate is set at 1000 Hz.

3.2.2 | Task preparation

Our experiment is based on a user behavior dataset collected from a one-month field study exploring the impact of search intent on user behavior and satisfaction in image search scenarios (Wu et al., 2019). The reason for choosing this dataset is that it records the real search logs of users and contains user search tasks in real environments, and thus it is more reliable and realistic than traditional lab study. Also, this dataset involves users' search intents, providing an opportunity to investigate factors that influence users' relevance judgments. This dataset has been used in several previous works (Xie et al., 2020; Xie, Mao, Liu, de Rijke, Shao, et al., 2019) on web image search scenarios.

From 555 search tasks and 2040 search queries in this dataset, 59 search tasks are carefully selected for the user experiments in our experiment, covering topics including science, sports, traveling, art, and so on. Pinkosova et al. (2023) suggest that, when participants indicate having SPK of the answer to a question, cognitive processing becomes easier. To eliminate this bias, our selection criteria include choosing unambiguous and easily understandable queries, while excluding ambiguous ones (e.g., "Shining") and relatively unknown ones (e.g., "SNH48"). Then, for the selected search task, a total of 724 images are crawled from Bing[3] image search engine as data to be labeled, with an average of 12.27 images per task. Examples of query descriptions are provided in Table 1.

In line with the approach adopted by Shao et al. (2019), we provide comprehensive instructions for each grade in the point-wise paradigm, as illustrated in Table 2. This step aims to minimize the subjectivity of assessors and enhance the consistency of scoring.

TABLE 1 Examples of query descriptions.

Query	Search intent description
Trigonometric functions	An image of trigonometric functions is needed for making slides
The Forbidden City	A friend is interested in knowing about the landmarks within the Forbidden City
Pearl necklace	I'm searching online for necklace styles because it's my mother's birthday on Saturday
Da Vinci	I have an assignment to search for Leonardo da Vinci's artworks

TABLE 2 Guidance instructions assigned to each grade for the experiment of point-wise paradigm (a higher score means higher relevance).

Score	Detailed guidance instructions
5	The objects and modifiers described in the requirements are perfectly matched in the given image
4	The object described in the requirement is matched in the image, but the modifiers do not fully match
3	The image partially meets the requirement (e.g., two or more objects are mentioned in the requirement, but only one appears in the image)
2	Only a small part of the image meets the requirements, but the main object described in the requirements is ambiguous in the image
1	The image does not meet the requirements at all

3.3 | Procedure

To protect the privacy and physical health of the subjects, our user study strictly adheres to the ethical procedures for the protection of human participants in research and is approved by the ethics committee of the School of Psychology at Tsinghua University.

As shown in Figure 1, our main experiment consists of two subtasks. Every participant is informed to participate in any number of subtasks voluntarily. It is important to note that the same dataset is used for both subtasks. For participants involved in both subtasks, we randomize the order of the two paradigms and ensure that the interval between their participation in the two experiments is at least 7 days. The 7-day interval is meant to prevent any potential bias caused by the memory effect (Fisher & Radvansky, 2018) so that we can collect annotations on the same query in different paradigms without letting the two subtasks affect each other.

Before the user study, participants are asked to complete an entry questionnaire to report demographic information and sign a consent about privacy security and personal information protection. They will be briefed about the main tasks and operation methods. They are also informed that they have the right to withdraw at any time during the study. Before the main trials, participants will undergo several training trials that resemble the main task. These training trials aim to familiarize participants with the overall process of the formal experiments. In the following sections, we detail descriptions of the two subtasks.

3.3.1 | Pair-wise relevance judgment

The green box in Figure 1 describes the process of pair-wise relevance judgment. Each participant needs to choose a random seed before the experiment to shuffle the order of the search tasks and the images. This randomization helps minimize potential biases and ensure a fair distribution of tasks and images across participants. The experimental platform follows a sequential and repetitive process, consisting of four steps: S1–S4, as illustrated in Figure 1 (S1). The experimental platform displays the current query description, that is, the query and search intent. Participants can proceed to the next stage by pressing the space key after understanding the task goal. (S2) A fixation cross is presented on the screen center to focus the field of vision so that attention can be drawn when the images appear. This fixation period lasts for 1000 ms. (S3) An image pair in the current task will be displayed successively in random order, and the display time of each image is 1500 ms. (S4) A 3-level weak pair-wise relevance judgment is to be made by participants through the keyboard. Figure 1 shows an example of pair-wise relevance judgment of the query "The Forbidden City."

3.3.2 | Point-wise relevance judgment

The blue box in Figure 1 describes the process of point-wise relevance judgment. The overall process follows a similar pattern as described above. To ensure randomness, a random seed is also selected at the beginning to shuffle the order of tasks and images. Steps S1 and S2 resemble that in the pair-wise relevance judgment process. (S3) A single image in the current task will be displayed, and the display time of each image is 1500 ms. (S4) A point-wise relevance judgment is to be made by participants through the keyboard. We use 5-level grading, and clear guidance instructions for each grade are provided in Table 2. Figure 1 shows an example of point-wise relevance judgment of the query "The Forbidden City."

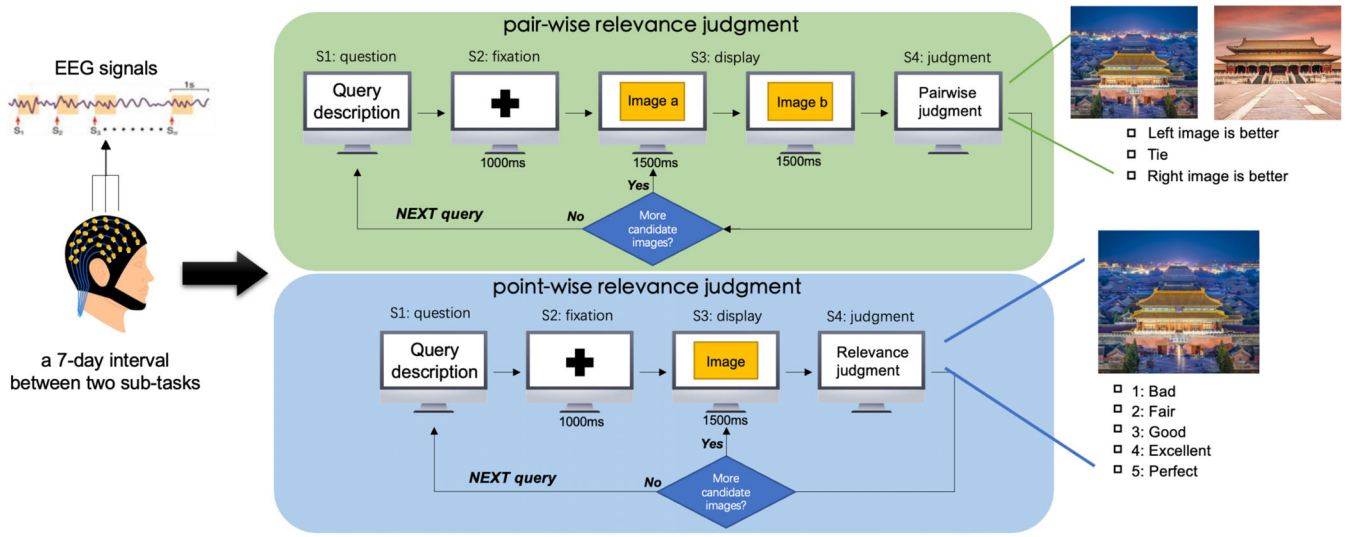


FIGURE 1 Structure of the main part of the two-subtask experiment, that is, pair-wise relevance judgment and point-wise relevance judgment. These subtasks are carried out in a randomized order and are spaced at least 7 days apart. (S1) Two subtasks begin with a query description on the screen. (S2) When participants press the space key, a fixation cross will be presented for 1000 ms. (S3) Then images will be displayed in the corresponding paradigm and every image will remain for a duration of 1500 ms. (S4) Finally, a relevance judgment will be made by participants through the keyboard. The program will return to S1 and display the next task description until there are no remaining tasks.

Throughout the main experiment, the participant's EEG signals will be captured and recorded during the whole process, together with the pre-coded triggers to locate time points of important events. In both paradigms, the program will return to step S1 and display the next task description, continuing this process until all the images in the current session have been presented. All images are divided equally into several sessions (9 in pair-wise and 7 in point-wise, which is determined in the pilot studies), and participants are allowed to take a rest between sessions.

3.4 | Pilot study

A pilot study can ensure the correctness of the overall experimental process and the reliability of the acquisition equipment. We conduct a pilot study on three participants whose data are not included in the final analysis to determine hyperparameters of the experimental presentation and design, such as the image presentation time, font size, number of sessions, and so on.

4 | RESULT ANALYSIS

This section analyzes the similarities and differences between the two paradigms through statistical methods and ERP analysis. By doing so, we aim to provide insights into the underlying neurological mechanisms associated with these paradigms. Based on the findings from our analysis, we discuss their implications to relevance judgment in IR scenarios.

4.1 | Statistics analysis

Among all participants, 15 individuals take part in the point-wise relevance judgment task, all of whom also participate in the pair-wise subtask. Additionally, four participants also take part in the pair-wise relevance judgment task only. However, one participant withdrew from the study prematurely, resulting in the absence of recorded data for that individual.

To facilitate subsequent ERP analysis and significance testing, we classified the EEG signals collected under both paradigms into two categories: **highRel** and **lowRel**. In the pair-wise paradigm, we focus on the second image in each pair and classify based on whether participants choose "**the second image is better**." In contrast, in the point-wise paradigm, we focus on each image and classify based on **participants' annotated relevance score**. Since our research objective is to compare the ordinal relationships of image relevance under two different paradigms, we excluded the neutral options (i.e., "Tie" in the pair-wise paradigm and the rating of 3 in the point-wise paradigm) in the ERP analysis. We report the distribution of different judgments per participant in Table 3.

TABLE 3 Statistics of the averaged number of judgments.

Relevance judgment	#Judgments
highRel (pair-wise)	170.8
lowRel (pair-wise)	175.5
highRel (point-wise)	285.3
lowRel (point-wise)	302.9

Note: In the pair-wise paradigm, "highRel"/"lowRel" means the second image has higher/lower relevance compared to the first image, while in the point-wise paradigm, they represent high-relevance score and low-relevance score, respectively.

In terms of time consumption, the average time spent in the judgment step of pair-wise relevance judgment (S4 in Figure 1) is 937 ms (SD = 102 ms), while the average time spent in the judgment step of point-wise relevance judgment (S4 in Figure 1) is 1207 ms (SD = 145 ms). This finding is in line with prior research observations that assessors tend to make quicker judgments in the pair-wise paradigm compared to the point-wise paradigm (Carterette et al., 2008; Xie et al., 2020).

4.2 | ERP methods

ERP refers to the steady voltage in the brain that is produced in response to a specific event or stimulus (Blackwood & Muir, 1990). Its advantage lies in its high time resolution, and the sequence formed by ERP peaks can accurately reflect the neural activity in the brain (Luck, 2014). ERP components are evoked amplitudes in different time windows, for example, N400 (negative waves within 400 ms) and P100, P300, P600 (positive waves within 100, 300, 600 ms). These commonly used ERP components capture distinct stages of neural processing, such as early sensory processing (P100), semantic processing (N400), attention allocation and decision-making (P300), and language syntax and grammar processing (P600). Previous studies have shown that ERP components also exhibit some fixed patterns in the procedure of relevance judgments (Pinkosova et al., 2020, 2022). We apply standard ERP analysis methods to the recorded data, including preprocessing, dividing time window, and identifying region of interest (ROI) (Luck, 2014).

4.2.1 | Preprocessing

Similar to previous work (Pinkosova et al., 2020; Ye et al., 2022), our preprocessing process includes: (1) Re-referencing the collected EEG data using the offline linked mastoids method (Yao et al., 2019). (2) Applying notch, low-pass, and high-pass filters to remove environmental noise, voltage drift, and high-frequency noise, respectively. (3) Employing the FASTER toolkit (Nolan et al., 2010) to eliminate bad channels and artificial components. (4) Extracting epochs of interest from the EEG signal sequences and calculating their average. The data epoch refers to the specific time window in which we expect to observe relevant brain responses. In our experiment, the data epoch spans from 200 ms before the stimulus (the presentation of images) to 1000 ms after the stimulus.

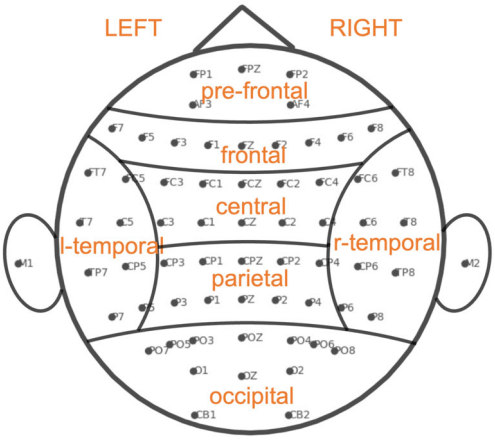


FIGURE 2 Seven brain regions according to their placement on the brain topography, that is, prefrontal, frontal, central, parietal, l-temporal, r-temporal, and occipital.

4.2.2 | Time windows

In order to separate different components in ERP, we split the extracted time interval into several time windows following the method proposed by Lehmann and Skrandies (1980). They recommend determining the components of evoked scalp potentials in terms of times of latency and topography. We calculate Global field power (GFP) between 50 and 950 ms, and split time windows by the peaks of GFP, which has been applied to existing ERP analyses researches (Ye et al., 2022).

4.2.3 | ROIs

Meanwhile, modern brain science research shows that different brain regions often have different functions. We divide the electrodes into seven brain regions according to their placement on the brain topography shown in Figure 2. We apply some statistical methods including ANOVA in a fixed time window for each brain region.

TABLE 4 Statistical significance differences of different time windows and ROIs.

Time windows and ROIs in pair-wise paradigm		
Time window	**ROI**	**Post hoc test**
50–200 ms	R-temporal	lowRel>highRel*
180–380 ms	Frontal	lowRel>highRel*
	Central	lowRel>highRel*
320–500 ms	Frontal	lowRel>highRel*
	R-temporal	lowRel>highRel**
550–650 ms	Central	lowRel>highRel*
	L-temporal	lowRel>highRel**
	R-temporal	lowRel>highRel*
	Parietal	lowRel>highRel*
Time windows and ROIs in point-wise paradigm		
Time window	**ROI**	**Post hoc test**
50–180 ms	Central	highRel>lowRel**
150–380 ms	Frontal	highRel>lowRel*
	Central	highRel>lowRel*
	Prefrontal	highRel>lowRel*
350–500 ms	L-temporal	highRel>lowRel*
	R-temporal	highRel>lowRel*
500–700 ms	L-temporal	highRel>lowRel**
	R-temporal	highRel>lowRel**
	Parietal	highRel>lowRel*

Note: Statistical significance at a level of $^{*}p < 0.05$, $^{**}p < 0.001$, respectively.

Mauchly's test is applied to verify the sphericity assumption and post hoc Bonferroni tests are employed to make pairwise comparisons between groups (Ye et al., 2022).

4.3 | ERP results

The statistical analysis results of time windows and ROIs are shown in Table 4. Based on the results, we exhibit the grand average ERP waveforms for each ERP component and corresponding ROIs in Figure 3. We will now provide explanations of the characteristics and potential functions associated with each ERP component.

4.3.1 | P100

P100 (presented in Figure 3a) is an early component in time window around 100 ms (50–200 ms for pair-wise and 50–180 ms for point-wise). We employ ANOVA statistical method and discover significant differences between the grand-averaged P100 component in r-temporal ($F[1,18] = 7.23, p < 0.05$) of pair-wise paradigm and in central ($F[1,14] = 20.96, p < 0.001$) of point-wise paradigm. Post hoc Bonferroni tests afterward reveal that P100 amplitude of "lowRel" ("highRel") is significantly higher than "highRel" ("lowRel") in the pair-wise (point-wise) paradigm with p-value <0.01 (0.001).

P100 component is considered to reflect the "cost of attention" (Luck, 2014), specifically, the initial ability and processing effort involved in recognizing relevant stimuli (Rutman et al., 2010). Higher amplitude in "highRel" of the point-wise paradigm suggests more early selective attention is allocated to highly relevant images in the process of initial visual field activation (Luck, 2014). On the other hand, in the pair-wise paradigm, we find the "lowRel" one evokes a greater amplitude of P100, which indicates the differences in attention distribution between the two paradigms.

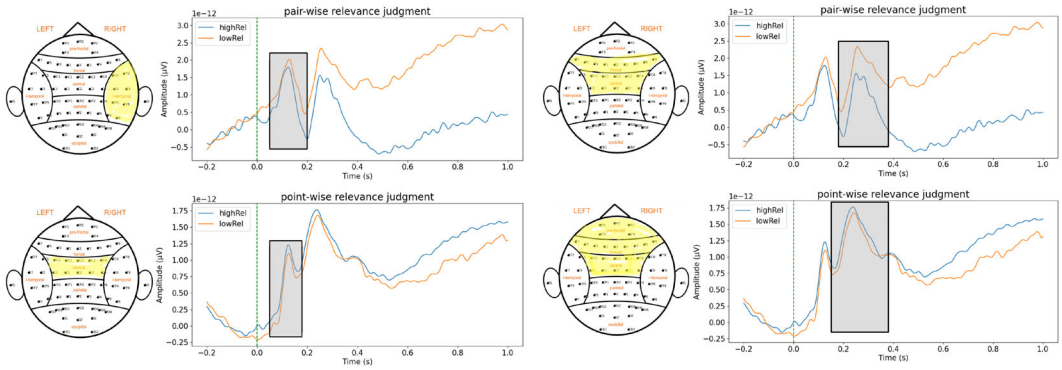
4.3.2 | P300

P300 (presented in Figure 3b) waveform is the dominant component in time window around 300 ms (180–380 ms for pair-wise and 150–380 ms for point-wise). ANOVA reveals the significant differences between grand-averaged P300 component in frontal ($F[1,18] = 10.14, p < 0.05$), central ($F[1,18] = 14.88, p < 0.05$) of pair-wise paradigm and in frontal ($F[1,14] = 12.34, p < 0.05$), central ($F[1,14] = 9.32, p < 0.05$), prefrontal ($F[1,14] = 10.86$, $p < 0.05$) of point-wise paradigm. Post hoc tests using the Bonferroni method also indicate that P300 amplitude of "lowRel" ("highRel") is significantly higher than "highRel" ("lowRel") in the pair-wise (point-wise) paradigm with p-value <0.001 (0.001).

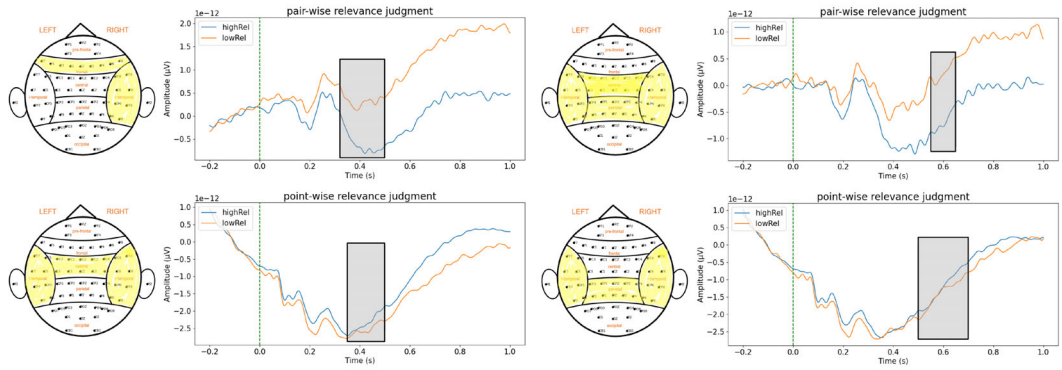
P300 component is considered as an endogenous potential in the process of decision making. It has been demonstrated that P300 is associated with brain activity related to the engagement of attention and the processing of novelty (Polich, 2007). In the point-wise paradigm, a higher amplitude of P300 indicates that more attention is allocated to highly relevant information, which also causes less cognitive load and memory load for assessors (Ahmed & de Fockert, 2012; Gray et al., 2004). But in the pair-wise paradigm, the "lowRel" one is easier to process and recognize since assessors will be primed on the first image, thus getting more attention.

4.3.3 | N400

N400 (presented in Figure 3c) waveform mainly appears in time window around 400 ms (320–500 ms for pair-wise



(a) *P100 component.* (b) *P300 component.*



(c) *N400 component.* (d) *P600 component.*

FIGURE 3 The grand average ERP waveforms for electrode configurations associated with ERP components, with the time interval of interest (highlighted in gray) and corresponding ROIs (highlighted in yellow). Refer to Figure 2 for detailed EEG channels information.

and 350–500 ms for point-wise). Through ANOVA, we find significant differences between grand-averaged N400 component in frontal ($F[1,18] = 8.09, p < 0.05$), central ($F[1,18] = 10.04, p < 0.05$), r-temporal ($F[1,18] = 22.16$, $p < 0.001$) of pair-wise paradigm and in l-temporal ($F[1,14] = 13.77, p < 0.05$), r-temporal ($F[1,14] = 12.38$, $p < 0.05$) of point-wise paradigm. Then we employ post hoc Bonferroni tests and find that N400 amplitude of "lowRel" ("highRel") is significantly higher than "highRel" ("lowRel") in the pair-wise (point-wise) paradigm with p-value <0.01 (0.001).

There have been some related studies on N400 showing that it is associated with uncertainty and incongruity of images and videos (Kim & Kim, 2019; Stuss et al., 1986). In the point-wise paradigm, reduced N400 is associated with lower relevant images which require more effort to infer a conclusion (Debruille, 2007). In the pair-wise paradigm, however, the "highRel" one has a lower N400 amplitude under the influence of the first image as a comparison.

4.3.4 | P600

P600 (presented in Figure 3d) component is evoked around 600 ms after the stimulus (550–650 ms for pair-wise and 500–700 ms for point-wise). Significant differences are found through ANOVA between grand-averaged P600 component in central ($F[1,18] = 11.40$, $p < 0.05$), l-temporal ($F[1,18] = 17.68, p < 0.001$), r-temporal ($F[1,18] = 9.07, p < 0.05$), parietal ($F[1,18] = 10.75, p < 0.05$) of pair-wise paradigm and in l-temporal ($F[1,14] = 20.91, p < 0.001$), r-temporal ($F[1,14] = 23.63$, $p < 0.001$), parietal ($F[1,14] = 12.88, p < 0.05$) of point-wise paradigm. Then we employ post hoc Bonferroni tests and find that P600 amplitude of "lowRel" ("highRel") is significantly higher than "highRel" ("lowRel") in the pair-wise (point-wise) paradigm with p-value <0.001 (0.001).

Previous researches find inferential processing (Burkhardt, 2006) will evoke P600 waveform. In IR scenarios, the link between higher relevance and P600

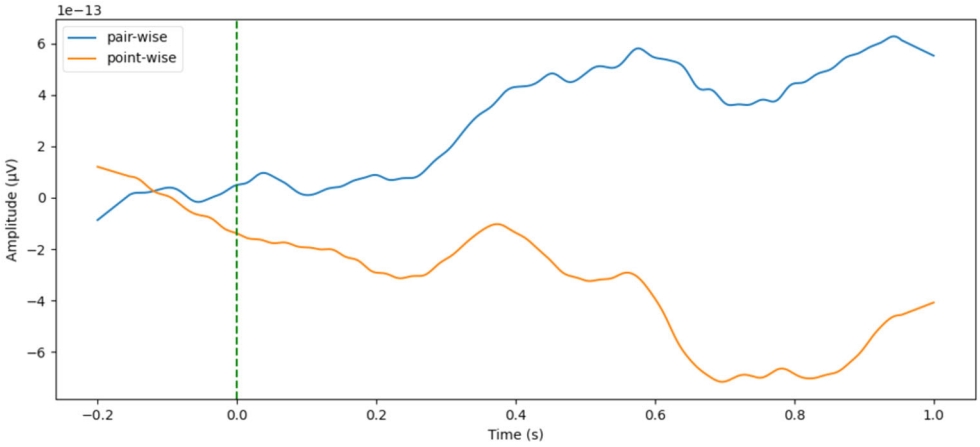


FIGURE 4 The difference waves in the central brain region of the pair-wise and point-wise paradigms ("lowRel"-"highRel").

amplitude is speculated to come from discourse memory in the brain (Yang et al., 2019), which is consistent with our experiment results. Pinkosova et al. (2020) indicate that the ease of the categorization process and more amount of information carried by the processed term (Kangassalo et al., 2019) is the reason why highly relevant images evoke higher P600 amplitude. The "highRel" one in the pair-wise paradigm is also of a similar nature for the same reasons, thus generating a greater P600 component.

4.4 | Discussion

The above analysis shows that the significant difference in each component of ERP is consistent with the overall trend difference between the two paradigms. All the EEG signal patterns analyzed of point-wise relevance judgment above are aligned with previous works on relevance judgment (Moshfeghi et al., 2016; Pinkosova et al., 2020, 2022). However, when we shift our focus to the pair-wise paradigm, we find that all conclusions are, in fact, the opposite (addressing RQ1).

To explore these different observations in the two paradigms, we further analyze their difference waves (Luck, 2014). Difference waves are calculated by subtracting the ERP waveforms of the two stimulus types (i.e., highRel and lowRel in our experiment) in the two paradigms. Difference waves are a commonly used approach to extract a purer ERP component with clearer psychological meaning. The curve after the Savitzky–Golay filter is shown in Figure 4. Difference waves technology can explain the process of mental activity more clearly and find out components that cannot be observed directly in the original waveform. Through the ANOVA method, we found that in each time window of the difference wave, there are significant differences between the two paradigms (all $p < 0.001$). This also confirms the interesting conclusion of ERP component analysis.

In other words, in the point-wise paradigm, people tend to pay more attention to images that are more satisfying to their needs, but in the pair-wise paradigm, people allocate more attention to relatively worse images (addressing RQ2).

One possible explanation for this observation is that in the point-wise relevance judgment, assessors may have a "golden standard" image in mind based on the task description and search intent. This ideal image meets all the given requirements and aligns with the assessor's personal aesthetic preferences. Consequently, highly relevant images that closely match this ideal standard will be evaluated more effortlessly, as previously discussed (Pinkosova et al., 2020, 2022). In contrast, during the pair-wise relevance judgment, assessors are exposed to an initial image before making their judgment. This initial image replaces the role of the "golden standard" in their mental representation. When encountering worse images, assessors produce the same neurological response as the point-wise paradigm, but better images instead get less attention and create a less memory load.

Furthermore, we examine the overall relationship between the two paradigms, and the grand average ERP waveforms are plotted in Figure 5. It is obvious that the amplitude of the ERP waveforms induced in the pair-wise relevance judgment is greater than that of the point-wise relevance judgment, which indicates that in the pair-wise paradigm, the assessors have less cognitive load and more focused attention allocation (Gray et al., 2004; Rutman et al., 2010). This is why assessors tend to respond faster in pair-wise paradigm (Carterette et al., 2008; Clarke et al., 2021; Xie et al., 2020). Then we consider the ERP of the first image in pair-wise relevance judgment together and notice that its trend is similar to the ERP curve of the point-wise paradigm (addressing

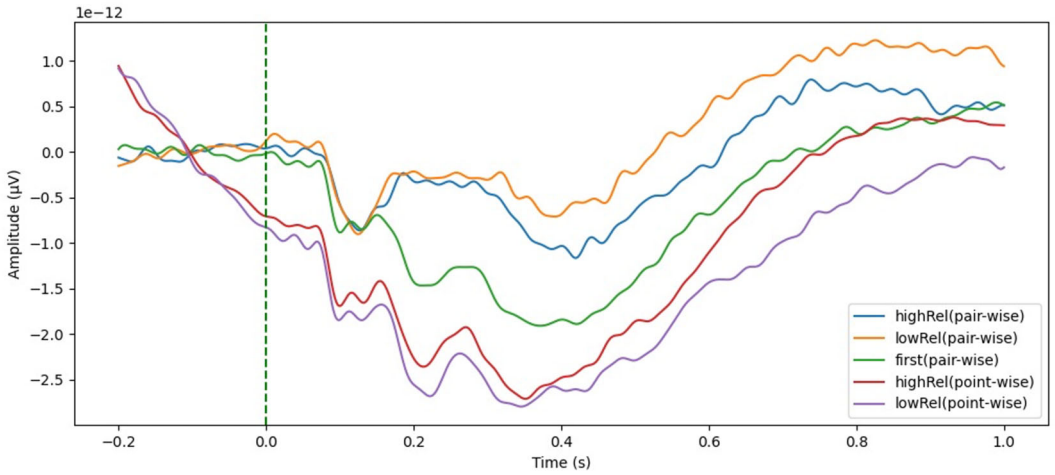


FIGURE 5 The grand average ERP waveforms in the central brain region of the two paradigms and the first image in pair-wise relevance judgment (green line).

RQ1). This implies that the assessors' judgment of the first item in the pair-wise relevance judgment has a similarity to the point-wise paradigm (addressing RQ2).

In conclusion, we analyze and compare the two paradigms from the underlying neurological mechanism, and provide detailed discussions. Our findings provide a deeper understanding of relevance judgments across different paradigms at the cognitive level, and illustrate insights for modern search techniques.

- **On search engine improvements**: When designing SERP, presenting search results in blocks or cards can encourage more pair-wise relevance judgments to alleviate the cognitive load of users, thereby facilitating easier decision-making. Users, unlike annotators, do not need to perform all pair-wise comparisons when browsing a grid-based SERPs, making it easier for them to focus on superior items over those low-scoring items that require more attention. Based on this, we suggest arranging low-quality items around prominently recommended items (such as sponsored products) to increase user attention and interest, thereby achieving the effect of implicit recommendation. Moreover, modern search engines' retrieval heavily relies on point-wise relevance assessment, where annotators devote more attention to higher-scoring items. Therefore, lower-scored items are not necessarily of low relevance. We suggest conducting additional pair-wise annotation tasks to make the annotation scores more accurate.
- **On annotation task design**: We observe a lower consistency in the point-wise relevance judgment in comparison with pair-wise relevance judgment due to variations in annotators' annotation standards. We propose pre-annotating a highly relevant item as a reference standard for annotators to establish a unified annotation standard. This step can reduce the cognitive load of annotators and improve the consistency of annotation scores. Furthermore, in pair-wise paradigms, annotators experience a higher cognitive load when judging the first item. To streamline the annotation process, fixing the reference point (first item) in the pair-wise paradigm can help reduce the cognitive burden and increase efficiency.

5 | PREDICTING RELEVANCE FROM BRAIN SIGNALS

Inspired by the concept of "brainsourcing" (i.e., utilize brain responses of a group of human contributors each performing a recognition task to determine classes of stimuli) (Davis III et al., 2020), we consider whether brain signals can be used as implicit feedback to predict annotators' judgments of preference.

5.1 | Experiments

Traditionally, two kinds of features are employed in the EEG prediction model, that is, event-related-potential-based features (ERPFs) and frequency-band-based features (FBFs) (Ye et al., 2022). ERPFs are time domain features in a specific short time window, especially where the ERP component is significantly different. FBFs are frequency domain features. It has been demonstrated that frequency information in different bands is associated with attentiveness (delta and beta; Harmony et al., 1996), cognitive performance (theta and alpha; Klimesch, 1999), and semantic violation (gamma;

TABLE 5 AUC of pair-wise relevance judgment prediction with brain signals of different brain regions in the within-subject experiments.

Brain region	LR	SVM	RF	GBDT
All	0.678	0.653	**0.710**	0.702
Frontal	0.629*	0.612*	0.689*	**0.697**
Central	**0.688**	**0.680**	0.617*	0.654*
Parietal	0.589*	0.607*	0.653*	0.684*
Temporal	0.676	0.637*	0.659*	0.641*
Occipital	0.587*	0.559*	0.561*	0.568*

Note: The bold values correspond to a p-value of 5%.
*indicates that the difference compared to the best-performing model is significant with p-value < 0.05.

TABLE 6 AUC of pair-wise relevance judgment prediction with brain signals of different brain regions in the cross-subject experiments.

Brain region	LR	SVM	RF	GBDT
All	0.535	0.537	0.553	0.551
Frontal	0.533	0.536	0.503*	0.521
Central	0.534	0.519	0.527	0.505*
Parietal	0.498*	0.511	0.513*	0.519
Temporal	0.510	0.490*	0.526	0.531
Occipital	0.501*	0.468*	0.494*	0.511*

*indicates that the difference compared to the best-performing model is significant with p-value < 0.05.

Penolazzi et al., 2009). We choose the latter in our experimental settings and calculate differential entropy from the frequency bands of delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–31 Hz), and gamma (31–81 Hz).

We focus on pair-wise relevance judgment in our experimental settings. The reason is that previous works have tried to use EEG signals as implicit feedback in the point-wise paradigm of binary classification and achieved decent results (Davis III et al., 2020, 2021; Gwizdka et al., 2017). The pair-wise relevance judgment can help the annotators to complete the judgment faster and more easily, thus has a specific application prospect and has not been studied as we know.

Considering that there are obviously individual differences in EEG signals, we first separately train and test a prediction model on the pair-wise relevance judgment data in different brain regions of each subject, namely the within-subject design. Then, we try to train a model on one subject and to predict the preference judgment of the others, which is called the cross-subject design.

The predicting models include Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Decision Tree (GBDT). Ten-fold cross-validation is employed to evaluate their performance.

The results of pair-wise relevance judgment prediction with brain signals in within-subject experiments are shown in Table 5. Among the prediction results of all models and brain region combinations, the best is the random forest model trained on all-electrode signals. When the frequency-domain features of all-electrode, frontal, central, and parietal are selected as the model input, the four models can all achieve decent performance, which is consistent with the statistical analysis of the ERP time windows and ROIs in the previous chapter. We notice that the performance on occipital is relatively poor. The occipital region is generally considered to undertake functions related to visual feedback. For example, make use of SSVEP for frequency prediction of visual stimuli to users (Chen et al., 2022). We explain that low-order vision-related functions are less associated with relevance judgments in general.

5.2 | Results and discussions

The results verify the feasibility of using brain signals as implicit feedback to predict relevance judgments. It also shows that this process can be completed in a very short time (within 950 ms in the pair-wise paradigm) using only part of the electrodes. With the help of brain signals, we can apply the previous methods combining pair-wise and point-wise relevance judgment with more direct feedback, thus ensuring the continuity of the annotators when using the system (addressing RQ3).

The results of pair-wise relevance judgment prediction with brain signals in the cross-subject scenarios are shown in Table 6. It can be seen that models trained on features of full-electrode still have the highest performance. However, compared with the prediction within-subject, the performance has dropped a lot even after having a wider range of training data.

The average AUC drops to around 0.5, which means models make a random classification. This is exactly the performance of individual differences in EEG signals. The problem of cross-subject prediction is always thorny in the field of EEG research. It is often difficult to find a general model to predict the performance of different subjects. Therefore, in practical applications, we can adopt algorithms that learn from transfer learning. The model is initially trained on large-scale network data, and before being applied to a specific subject, it is first fine-tuned according to their specific EEG characteristics to better adapt.

6 | CONCLUSION

In this paper, we conduct a well-designed lab-based user study to compare point-wise and pair-wise relevance judgment in image search scenarios from a neuroscience perspective. Through ERP analysis, we mainly conclude the following insightful findings: (1) Significant differences exist in N400 and P100, P300, P600 components between the two paradigms. We suggest that people tend to judge how good an item is in the point-wise paradigm while trying to figure out which is worse in the pair-wise paradigm. (2) The users' brain activities during relevance judgment of the first item in the pair-wise paradigm are similar to that in the point-wise paradigm. (3) Less cognitive load is the reason helps assessors to make judgments faster and more easily in the pair-wise paradigm. Based on these findings, we explore the potential implications for future IR tasks related to relevance judgment. Furthermore, we verify the feasibility of employing EEG signals for predicting relevance judgment within subjects and show that EEG signals can serve as implicit user feedback which would be helpful in IR.

The limitations of our work that may guide future works include: (1) Our study is limited to a lab-based environment under our experimental paradigm, that is, 3-level weak preference judgment and 5-level point-wise relevance judgment. Although we classified the stimuli into two categories to make our findings universal, the impact of graded relevance judgment paradigms on the results remains to be studied. (2) Our experiment was conducted in the context of image search, however, our findings align with previous research conclusions, suggesting that the experimental paradigm we employed can be extended to relevance assessment studies in multimodal scenarios. This direction of research represents a potential avenue for future investigations. (3) We only validate the feasibility of using brain signals as implicit feedback to predict user relevance judgments and the accuracy of prediction results still has room for improvement. Future works can explore the extraction and selection of brain signal features and design more sophisticated models that can achieve better performance.

ORCID

Shuqi Zhu https://orcid.org/0009-0001-7167-282X
Xiaohui Xie https://orcid.org/0000-0001-9413-4461
Ziyi Ye https://orcid.org/0000-0002-5622-0235
Qingyao Ai https://orcid.org/0000-0002-5030-709X
Yiqun Liu https://orcid.org/0000-0002-0140-4512

ENDNOTES

[1] Note that, for the sake of clarity and maintaining consistency in terminology, we refer to the two paradigms as point-wise and pair-wise relevance judgment throughout the rest of this paper.

[2] https://github.com/Promise-Z5Q2SQ/Pointwise-Pairwise-EEG.

[3] www.bing.com.

REFERENCES

Ahmed, L., & de Fockert, J. W. (2012). Working memory load can both improve and impair selective attention: Evidence from the navon paradigm. *Attention, Perception, & Psychophysics*, *74*, 1397–1405.

Allegretti, M., Moshfeghi, Y., Hadjigeorgieva, M., Pollick, F. E., Jose, J. M., & Pasi, G. (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* (pp. 719–722). ACM.

Arabzadeh, N., Kmet, O., Carterette, B., Clarke, C. L., Hauff, C., & Chandar, P. (2023). A is for adele: An offline evaluation metric for instant search. In *Proceedings of the 2023 ACM SIGIR international conference on theory of information retrieval* (pp. 3–12). ACM.

Bah, A., Chandar, P., & Carterette, B. (2015). Document comprehensiveness and user preferences in novelty search tasks. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval* (pp. 735–738). ACM.

Blackwood, D., & Muir, W. J. (1990). Cognitive brain potentials and their application. *British Journal of Psychiatry*, *157*(S9), 96–101.

Burkhardt, P. (2006). Inferential bridging relations reveal distinct neural mechanisms: Evidence from event-related brain potentials. *Brain and Language*, *98*(2), 159–168.

Carterette, B., Bennett, P. N., Chickering, D. M., & Dumais, S. T. (2008). Here or there: Preference judgments for relevance. In *ECIR'08: Proceedings of the IR research, 30th European conference on advances in information retrieval* (pp. 16–27). Springer.

Chen, X., Ye, Z., Xie, X., Liu, Y., Gao, X., Su, W., Zhu, S., Sun, Y., Zhang, M., & Ma, S. (2022). Web search via an efficient and effective brain-machine interface. In *Proceedings of the fifteenth ACM international conference on web search and data mining*. ACM.

Chu, Z., Mao, J., Zhang, F., Liu, Y., Sakai, T., Zhang, M., & Ma, S. (2021). Evaluating relevance judgments with pairwise discriminative power. In *Proceedings of the 30th ACM international conference on information & knowledge management* (pp. 261–270). ACM.

Clarke, C. L., Craswell, N., & Soboroff, I. (2004). Overview of the trec 2004 terabyte track. *TREC*, *4*, 74.

Clarke, C. L., Vtyurina, A., & Smucker, M. D. (2021). Assessing top-preferences. *ACM Transactions on Information Systems*, *39*(3), 1–21.

Cleverdon, C. (1967). The cranfield tests on index language devices. *ASLIB Proceedings*, *19*(6), 173–194.

Collins-Thompson, K., Macdonald, C., Bennett, P., Diaz, F., & Voorhees, E. M. (2015). *Trec 2014 web track overview* (Technical Report). University of Michigan.

Davis, K. M., III, Kangassalo, L., Spapé, M., & Ruotsalo, T. (2020). Brainsourcing: Crowdsourcing recognition tasks via collaborative brain-computer interfacing. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–14). ACM.

Davis, K. M., III, Spapé, M., & Ruotsalo, T. (2021). Collaborative filtering with preferences inferred from brain signals. In *Proceedings of the web conference* (pp. 602–611). ACM.

Debruille, J. B. (2007). The n400 potential could index a semantic inhibition. *Brain Research Reviews*, *56*(2), 472–477.

Eugster, M. J., Ruotsalo, T., Spapé, M. M., Kosunen, I., Barral, O., Ravaja, N., Jacucci, G., & Kaski, S. (2014). Predicting term-relevance from brain signals. In *Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval* (pp. 425–434). ACM.

Fisher, J. S., & Radvansky, G. A. (2018). Patterns of forgetting. *Journal of Memory and Language*, *102*, 130–141.

Geng, B., Yang, L., Xu, C., Hua, X.-S., & Li, S. (2011). The role of attractiveness in web image search. In *Proceedings of the 19th ACM international conference on multimedia* (pp. 63–72). ACM.

Gray, H. M., Ambady, N., Lowenthal, W. T., & Deldin, P. (2004). P300 as an index of attention to self-relevant stimuli. *Journal of Experimental Social Psychology*, *40*(2), 216–224.

Gwizdka, J., Hosseini, R., Cole, M., & Wang, S. (2017). Temporal dynamics of eye-tracking and eeg during reading and relevance decisions. *Journal of the Association for Information Science and Technology*, *68*(10), 2299–2312.

Harmony, T., Fernández, T., Silva, J., Bernal, J., Díaz-Comas, L., Reyes, A., Marosi, E., Rodríguez, M., & Rodríguez, M. (1996). Eeg delta activity: An indicator of attention to internal processing during performance of mental tasks. *International Journal of Psychophysiology*, *24*(1–2), 161–171.

Hui, K., & Berberich, K. (2017). Low-cost preference judgment via ties. In *Advances in information retrieval: 39th European conference on IR research, ECIR 2017, Aberdeen, UK, April 8–13, 2017, proceedings* (pp. 626–632). Springer.

Jain, V., & Varma, M. (2011). Learning to re-rank: Query-dependent image re-ranking using click data. In *Proceedings of the 20th international conference on world wide web* (pp. 277–286). ACM.

Kangassalo, L., Spapé, M., Jacucci, G., & Ruotsalo, T. (2019). Why do users issue good queries? Neural correlates of term specificity. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval* (pp. 375–384). ACM.

Kim, H. H., & Kim, Y. H. (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), 931–941.

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis. *Brain Research Reviews*, *29*(2–3), 169–195.

Koh, J. Y., Fried, D., & Salakhutdinov, R. R. (2024). Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, *36*, 21487–21506.

Lefortier, D., Serdyukov, P., & De Rijke, M. (2014). Online exploration for detecting shifts in fresh intent. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management* (pp. 589–598). ACM.

Lehmann, D., & Skrandies, W. (1980). Reference-free identification of components of checkerboard-evoked multichannel potential fields. *Electroencephalography and Clinical Neurophysiology*, *48*(6), 609–621.

Liu, Y., Mao, J., Xie, X., Zhang, M., & Ma, S. (2021). Challenges in designing a brain-machine search interface. *ACM SIGIR Forum*, *54*(2), 1–13.

Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT Press.

Luck, S. J., Woodman, G. F., & Vogel, E. K. (2000). Event-related potential studies of attention. *Trends in Cognitive Sciences*, *4*(11), 432–440.

Luo, C., Sakai, T., Liu, Y., Dou, Z., Xiong, C., & Xu, J. (2017). *Overview of the NTCIR-13 we want web task*. NTCIR.

Michalkova, D., Rodriguez, M. P., & Moshfeghi, Y. (2024). Understanding feeling-of-knowing in information search: An EEG study. *ACM Transactions on Information Systems*, *42*(3), 1–30.

Moshfeghi, Y., & Jose, J. M. (2013). An effective implicit relevance feedback technique using affective, physiological and behavioural features. In *Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval* (pp. 133–142). ACM.

Moshfeghi, Y., Triantafillou, P., & Pollick, F. E. (2016). Understanding information need: An FMRI study. In *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval* (pp. 335–344). ACM.

Nolan, H., Whelan, R., & Reilly, R. B. (2010). Faster: Fully automated statistical thresholding for eeg artifact rejection. *Journal of Neuroscience Methods*, *192*(1), 152–162.

Penolazzi, B., Angrilli, A., & Job, R. (2009). Gamma EEG activity induced by semantic violation during sentence reading. *Neuroscience Letters*, *465*(1), 74–78.

Pinkosova, Z., McGeown, W., & Moshfeghi, Y. (2022). Revisiting neurological aspects of relevance: An EEG study. In *Machine learning, optimization, and data science* (pp. 549–563). Springer.

Pinkosova, Z., McGeown, W. J., & Moshfeghi, Y. (2020). The cortical activity of graded relevance. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 299–308). ACM.

Pinkosova, Z., McGeown, W. J., & Moshfeghi, Y. (2023). Moderating effects of self-perceived knowledge in a relevance assessment task: An eeg study. *Computers in Human Behavior Reports*, *11*, 100295.

Polich, J. (2007). Updating p300: An integrative theory of p3a and p3b. *Clinical Neurophysiology*, *118*(10), 2128–2148.

Radinsky, K., & Ailon, N. (2011). Ranking from pairs and triplets: Information quality, evaluation methods and query complexity. In *Proceedings of the fourth ACM international conference on web search and data mining* (pp. 105–114). ACM.

Roitero, K., Maddalena, E., Demartini, G., & Mizzaro, S. (2018). On fine-grained relevance scales. In *The 41st international ACM SIGIR conference on research & development in information retrieval* (pp. 675–684). ACM.

Rutman, A. M., Clapp, W. C., Chadick, J. Z., & Gazzaley, A. (2010). Early top–down control of visual processing predicts working memory performance. *Journal of Cognitive Neuroscience*, *22*(6), 1224–1234.

Sang, J., Xu, C., & Lu, D. (2011). Learn to personalized image search from the photo sharing websites. *IEEE Transactions on Multimedia*, *14*(4), 963–974.

Shao, Y., Liu, Y., Zhang, F., Zhang, M., & Ma, S. (2019). On annotation methodologies for image search evaluation. *ACM Transactions on Information Systems*, *37*(3), 1–32.

Stuss, D. T., Picton, T., & Cerri, A. (1986). Searching for the names of pictures: An event-related potential study. *Psychophysiology*, *23*(2), 215–223.

White, R. W., Ruthven, I., & Jose, J. M. (2002). *The use of implicit evidence for relevance feedback in web retrieval*. Springer.

Wu, Z., Liu, Y., Zhang, Q., Wu, K., Zhang, M., & Ma, S. (2019). The influence of image search intents on user behavior and satisfaction. In *Proceedings of the twelfth ACM international conference on web search and data mining* (pp. 645–653). ACM.

Wu, Z., Mao, J., Liu, Y., Zhang, M., & Ma, S. (2020). Investigating reading behavior in fine-grained relevance judgment. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 1889–1892). ACM.

Xie, X., Mao, J., de Rijke, M., Zhang, R., Zhang, M., & Ma, S. (2018). Constructing an interaction behavior model for web image search. In *The 41st international ACM SIGIR conference on research & development in information retrieval* (pp. 425–434). ACM.

Xie, X., Mao, J., Liu, Y., de Rijke, M., Ai, Q., Huang, Y., Zhang, M., & Ma, S. (2019). Improving web image search with contextual information. In *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 1683–1692). ACM.

Xie, X., Mao, J., Liu, Y., de Rijke, M., Chen, H., Zhang, M., & Ma, S. (2020). Preference-based evaluation metrics for web image search. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 369–378). ACM.

Xie, X., Mao, J., Liu, Y., de Rijke, M., Shao, Y., Ye, Z., Zhang, M., & Ma, S. (2019). Grid-based evaluation metrics for web image search. In *The world wide web conference* (pp. 2103–2114). ACM.

Yan, X., Luo, C., Clarke, C. L., Craswell, N., Voorhees, E. M., & Castells, P. (2022). Human preferences as dueling bandits. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval* (pp. 567–577). ACM.

Yang, H., Laforge, G., Stojanoski, B., Nichols, E. S., McRae, K., & Köhler, S. (2019). Late positive complex in event-related potentials tracks memory signals when they are decision relevant. *Scientific Reports*, *9*(1), 1–15.

Yang, X., Zhang, Y., Yao, T., Ngo, C.-W., & Mei, T. (2015). Click-boosting multi-modality graph-based reranking for image search. *Multimedia Systems*, *21*, 217–227.

Yang, Z., Moffat, A., & Turpin, A. (2018). Pairwise crowd judgments: Preference, absolute, and ratio. In *Proceedings of the 23rd Australasian document computing symposium* (pp. 1–8). ACM.

Yao, D., Qin, Y., Hu, S., Dong, L., Bringas Vega, M. L., & Valdés Sosa, P. A. (2019). Which reference should we use for EEG and ERP practice? *Brain Topography*, *32*(4), 530–549.

Ye, Z., Xie, X., Liu, Y., Wang, Z., Chen, X., Zhang, M., & Ma, S. (2022). Towards a better understanding of human reading comprehension with brain signals. In *Proceedings of the ACM web conference* (Vol. 2022, pp. 380–391). ACM.

Zhang, P.-F., Bai, G., Huang, Z., & Xu, X.-S. (2022). Machine unlearning for image retrieval: A generative scrubbing approach. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 237–245). ACM.

How to cite this article: Zhu, S., Xie, X., Ye, Z., Ai, Q., & Liu, Y. (2024). Comparing point-wise and pair-wise relevance judgment with brain signals. *Journal of the Association for Information Science and Technology*, 1–15. https://doi.org/10.1002/asi.24936