Temporal Dynamics of Eye-Tracking and EEG During Reading and Relevance Decisions

Jacek Gwizdka
School of Information, University of Texas at Austin, Austin, TX, 78701, USA. E-mail: jasist@gwizdka.com

Rahilsadat Hosseini
Department of Industrial & Manufacturing Systems Engineering, University of Texas at Arlington, Arlington, TX 76019, USA. E-mail: rahilsadat.hosseini@mavs.uta.edu

Michael Cole
LexisNexis, New York, NY, USA. E-mail: michael.cole@lexisnexis.com

Shouyi Wang
Department of Industrial & Manufacturing Systems Engineering, University of Texas at Arlington, Arlington, TX 76019, USA. E-mail: shouyiw@uta.edu

Assessment of text relevance is an important aspect of human–information interaction. For many search sessions it is essential to achieving the task goal. This work investigates text relevance decision dynamics in a question-answering task by direct measurement of eye movement using eye-tracking and brain activity using electroencephalography EEG. The EEG measurements are correlated with the user's goal-directed attention allocation revealed by their eye movements. In a within-subject lab experiment ($N = 24$), participants read short news stories of varied relevance. Eye movement and EEG features were calculated in three epochs of reading each news story (early, middle, final) and for periods where relevant words were read. Perceived relevance classification models were learned for each epoch. The results show reading epochs where relevant words were processed could be distinguished from other epochs. The classification models show increasing divergence in processing relevant vs. irrelevant documents after the initial epoch. This suggests differences in cognitive processes used to assess texts of varied relevance levels and provides evidence for the potential to detect these differences in information search sessions using eye tracking and EEG.

Introduction

Relevance is a central construct for information search and retrieval (IS&R; Borlund, 2003; Hjørland, 2010; Saracevic, 1975, 2007). We are interested in the human process of making relevance judgments (RJs). Although users can provide some reasons for RJs, the act of judging is opaque. A consequence is poor understanding of the factors affecting RJs (Huang & Soergel, 2013). Understanding RJ cognitive processes is a knowledge gap at the foundations of IS&R. This research seeks to contribute to bridging this gap. We use neurophysiological (NP) instruments to capture psychophysiological signals while a user is making RJs while engaged in information search.

Relevance research is extensive, ranging from theoretical (Hjørland, 2010; Huang & Soergel, 2013; Saracevic, 2007), to behavioral studies (Barry, 1994; Fitzgerald & Galloway, 2001; Taylor, 2012), and to applied and system-oriented evaluation studies (Lesk & Salton, 1968; Ruthven, 2014). RJs have a strong subjective component, yet there is surprisingly little work on internal psychological factors. Theoretical work exists (Wilson & Sperber, 2002), but empirical research addressing RJ cognitive and affective processes is quite recent (Allegretti et al., 2015; Buscher, Dengel, Biedert, & Elst, 2012; Moshfeghi, Pinto, Pollick, & Jose, 2013). Improved understanding of cognitive processes that drive IS&R may eventually allow for causal accounts of user information search behaviors.

Direct elicitation of user RJs poses challenges. Most users will not make the effort to provide explicit relevance feedback (Back & Oppenheim, 2001), and reporting RJs can reflect biases. Inferring relevance implicitly from user actions is attractive (White, Ruthven, & Jose, 2002).

Preprint Manuscripts to ASIS&T 2017

This is not for mass dissemination just for students and colleagues

However, many available signals such as dwell time (Kelly & Belkin, 2004), click-throughs (Jung, Herlocker, & Webster, 2007), mouse movement (Smucker, Guo, & Toulis, 2014), and other interactions (Kelly & Belkin, 2001), often provide ambiguous evidence.

Inferring relevance from NP signals holds promise for nonintrusive detection in natural settings. Here we investigate relevance detection using eye tracking and electroencephalography (EEG). Eye movements are cognitively controlled (Just & Carpenter, 1987). Brain activity is detectable using EEG. Visual attention to text is a central process for information acquisition and judging relevance. Word understanding can be distinguished from other higher-level cognitive process in judging text relevance (Park & Reder, 2004).

This work concerns text RJs in the cognitive activities of reading and judging news stories in a question-answering task. We explore activity correlations using EEG and eye tracking (EYE) during text processing of relevant words and in three time epochs. For each epoch, EEG- and EYE-based classification models of perceived relevance are learned. Epoch model performance is compared to explore text processing differences when relevant text is recognized. Model composition and performance differences can provide insights into the cognitive processes of text RJ.

We have previously reported EYE data analysis for whole documents (Gwizdka, 2014). To investigate user–document interaction, we now present higher-resolution analysis (1s, 2s). The main contribution is to show that there are increasing differences in EYE and EEG data between the relevant and irrelevant documents as text documents are read. Another contribution is to demonstrate a low-cost EEG device can be useful to infer RJs.

Related Work

Obtaining implicit relevance measures from user behaviors is an important research area in information retrieval (IR). Here we focus on relevance-related work using eye movement and EEG NP methods, which demonstrates the plausibility of monitoring brain activity and related NP signals to detect aspects of relevance processes.

RJ is a user assessment about the usefulness of a text to their task. It is produced by a complex cognitive process. IR research shows external user behaviors are usefully, but imperfectly, correlated with relevance assessments. Implicit relevance indicators include search results click-through (Dupret & Liao, 2010), document dwell time (Kellar, 2004; Kelly & Belkin, 2004; Liu & Belkin, 2010; Liu, Liu, & Belkin, 2014; White & Kelly, 2006), mouse movements (Cooke, 2006; Guo & Agichtein, 2010; Huang, White, & Buscher, 2012; Rodden, Fu, Aula, & Spiro, 2008; Smucker et al., 2014), and text selection actions (White & Buscher, 2012).

Eye-Tracking

Eye-tracking is a familiar tool for IR relevance investigations. Ajanki, Hardoon, Kaski, Puolamäki, and Shawe-Taylor (2009) used eye-movement-based features to select additional query terms in an implicit relevance feedback system. Regressions and first fixation were most useful. In a think-aloud study by Balatsoukas and Ruthven (2012), users made more frequent and longer fixations on nonrelevant document surrogates. However, research avoiding the cognitive process of think-aloud (Gwizdka, 2014; Villa & Halvey, 2013) shows not-relevant documents impose the lowest mental load.

Buscher et al. (2012) found the strongest indicator in text passage relevance was length of text read, whereas fixation duration was uncorrelated. Simola, Salojärvi, and Kojo (2008) improved search task performance using eye-movement features. Oliveira, Aula, and Russell (2009) investigated pupil dilation (PD) and showed relevant images led to increased PD.

Marcos, Gavin, and Arapakis (2015) studied eye and mouse movement behaviors on SERP snippets incorporating images, multimedia, and text. They developed measures of noticeability and interest using fixations, and conversion (RJs) using click-through.

EEG

IS&R has little EEG work as compared to extensive cognitive psychology research, where visual search and target differentiation is most germane to IR research. Eye-fixation-related potentials (EFRPs) combines event-related potentials (ERPs) based on stimulus EEG signals and eye fixations to define epoch timing. Brouwer, Reuderink, Vincent, van Gerven, and van Erp (2013) showed the use of ERPs to detect designated visual targets (accuracy 0.62). Healy and Smeaton (2011) extracted a P300 ERP that could identify cases where the visual target was disclosed beforehand. They inferred an EFPR from electro-oculographic EEG artifacts. EEG activity for nonprimed visual target recognition, which required participant judgment, began at ∼250 ms and persisted for 500–1,000 ms.

EEG IS&R work has investigated word processing EEG signals. Frey et al. (2013) investigated short text search (∼5 lines, 30 words total). They found post-RJ brain wave differences in processing relevant and irrelevant words that persisted for one word after a relevant word (∼260–320 ms) and two words after an irrelevant word (∼500–530 ms). Eugster et al. (2014) demonstrated frequency spectrum and ERPs (450–747 ms) could distinguish relevant and irrelevant terms (0.67 accuracy). Allegretti et al. (2015) showed image RJ brain activity peaks ∼500–800 ms after onset with evidence for cognitive processes during 300–500–800 ms.

EEG data, facial expressions, and eye gaze were used to model image tag relevance (Soleymani, Kaltwang, & Pantic, 2013). They found that tag relevance detection using eye gaze alone performed better at the top of rankings and over the entire list.

Inexpensive EEG Devices

We used an inexpensive EEG device (Emotiv EPOC) that, although less sensitive than medical-grade devices, has

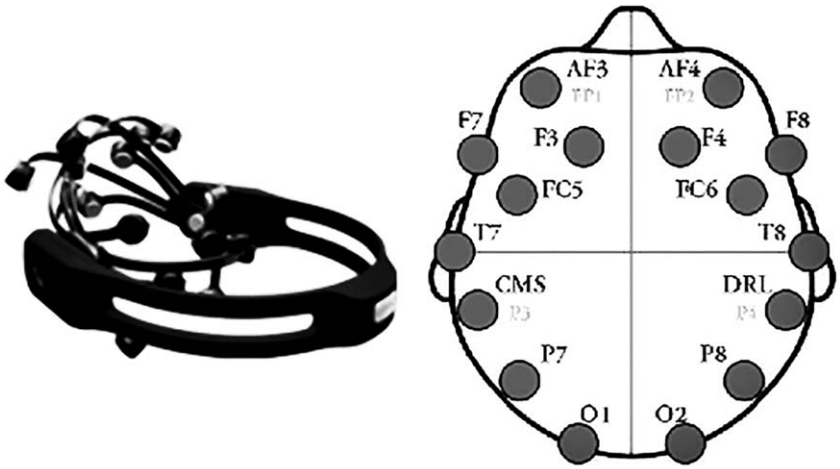

FIG. 1. Emotiv EEG headset with electrode positions in 10/20 system.

been used successfully in research (Abbott & Faisal, 2012; Bobrov, Frolov, Cantor, Bakhnyan, & Zhavoronkov, 2011; Khushaba et al., 2012; Ramirez, Palencia-Lefler, Giraldo, & Vamvakousis, 2015; Wang, Gwizdka, & Chaovalitwongse, 2015). ERP analysis limitations exist (Duvinage et al. 2013), but Badcock et al. (2013) reported successful ERP analysis using the EPOC. We use it for power spectrum analysis following other researchers, including investigation of cognitive workload (Wang et al., 2015), graph understanding (Anderson et al., 2011), and silent reading (Knoll et al., 2011). Another application has been classification in a target pointing and selection task (Kim, Kim, & Jo, 2015). Bobrov et al. (2011) found EPOC classification performance was comparable to a medical-grade device (BrainProducts ActiCap, Gilching, Germany) in a task with recognition of two image types (face or house) and a relaxation state (3-class accuracy [EPOC ActiCap]: overall 0.48 vs. 0.54; best 0.62 vs. 0.68; Bobrov et al., 2011).

Summary

Research has established that text relevance affects how the text is read and images viewed. These differences are reflected in EYE and EEG measures. NP methods using brain waves, eye movement, and PD dynamics can be used to study changing mental states in information search and detection of RJs. The reported EEG RJ studies employed medical-grade EEG devices. To our knowledge, no prior RJ research has employed low-cost EEG devices. Furthermore, those studies are limited to words, very short texts ($\leq$ 30 words), or images. Our work seeks to gain better understanding of RJs and dynamics of processing texts in significantly longer text passages.

Method

Research questions:

RQ1. Does the text document relevance affect person eye-movement and brain activity, as measured by EEG, differently at early, middle, late stages of reading?

RQ2. Can periods of reading relevant words be distinguished from periods of reading irrelevant documents using eye-tracking and EEG data?

RQ3. Can text document relevance be plausibly inferred from EEG signals obtained from a low-cost device alone and in combination with EYE data?

Experiment/Participants

We conducted an Institutional Review Board (IRB)-approved lab experiment in which undergraduate and graduate student participants ($N = 24$; nine females; native English speakers; normal or corrected vision) were asked to find information in short news story texts. Each participant received $25.

Apparatus

A 17″ Tobii T-60 eye-tracker (1280×1024) and an Emotiv EPOC EEG wireless headset (Figure 1) producing 2,048 samples/s (internally downsampled to 128 samples/s) using 14 channels, with positioning using the international 10/20 EEG format (Figure 1). The Emotiv EPOC device wirelessly captures EEG signals. However, in practice the acquired signals may be a mixture of electrical signals due to brain activity (EEG), eye movement (EOG), and other signals related to, for example, facial muscle activity. The latter two types of signals may be captured by the Emotiv system due to several prefrontal locations of electrodes (AF3 and AF4). Throughout the article, we use the term *EEG signals* to refer to the signals acquired by the Emotiv system, keeping in mind that sources of these signals may not be limited to brain activity. Participants were seated ∼65–75 cm from the monitor under fluorescent ceiling lights. Text lines spanned ∼0.9° of visual angle (32 pixels; 19pt Verdana font) and were displayed in a 1020×900 region to optimize eye-tracker accuracy.

Procedure

We conducted a within-subject experiment where each participant performed the following two types of tasks: (a)

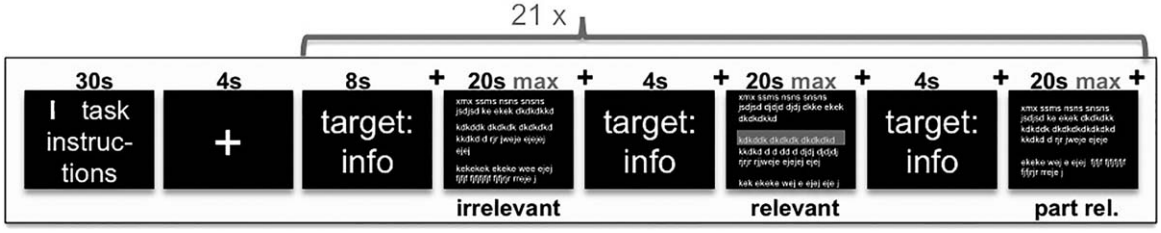

FIG. 2. Experimental design (task IS).

target word search (WS task) and (b) a simple question/answer information search (IS task). The experiment started with a WS and IS training task and the experiment session lasted about 75 minutes. WS and IS tasks were presented in balanced order. For this paper's research questions, only the IS task results are germane. The IS task is a simulated search task with human RJ. The question can be considered a simulated query with three news story texts as the top three search results. Overall, there were 21 questions followed by three short news stories each (the trials) and lasted a mean of 38 minutes ($SD = 2.5$). Figure 2 shows the stimulus sequence and timing for a single IS QA instance. The text documents were selected to cover a diversity of news topics. The experiment was presented using a fully automated process with controlled timing except for the text display duration, which ended when the participant responded with an RJ.

The IS task block started with general task instructions (30 seconds). The task goal was to find factual information in news stories that provided an answer to a question. The information target was presented as a question (8 seconds). Then a fixation cross appeared for 4 seconds to center the participant's eye gaze. Then the first news story was shown, followed by another fixation cross (4 seconds). To remind participants of the current question, it was repeated briefly (4 seconds) before the second and third document (Figure 2 "target info"). Fixation crosses were inserted between each of the news stories and the question reminders. Participants were asked to decide if the document contained an answer to the question or not and so a binary scale was used. The questions were factual in nature, for example: "Which Russian fleet was submarine Kursk part of?" News-story texts were presented for up to 20 seconds or until the participant pressed a key. This time limit was learned in pilot testing to ensure that participants could comfortably read the story and make a decision. The questions were presented in randomized order. Within a question block there were three trials; each trial was a news story text that was either:

- *Irrelevant* (I): a news-story on a topic different from the question,
- *Topical*, or partially relevant (T): a news-story on the question's topic, but not containing the question answer,
- *Relevant* (R): a news-story containing an answer to the question.

The new-stories were rotated within a question block in a pseudorandom manner using the following process. Each block contained one relevant document (R) and a combination of topical (T) or irrelevant (I) documents. Thus, within each question block there were the following three possible combinations of relevance levels: (a) RTT, (b) RTI, (c) RII. These combinations can be permuted in three, six, and three ways, respectively, yielding 12 orders: for RTT combination: RTT, TRT, TTR, for RTI: RTI, RIT, IRT, ITR, TRI, TIR, and for RII: RII, IRI, IIR. These 12 orders were used to create a sequence of 21 question blocks (nine of them were repeated twice; $12 + 9 = 21$). The sequence of 21 question blocks was randomized. We followed this procedure to avoid exposing participants to similar orders of document relevance from which they could plausibly learn patterns and try to guess relevance levels. A participant saw each document exactly once. It is important to note that participants were told that in any block some, all, or none of the stories could be relevant. This was done to enhance the ability to treat each text as involving an independent RJ.

The 21 IS task question blocks were further divided into "overt" and "covert" blocks. The overt block contained 14 question blocks (with $14*3 = 42$ trials). In this block participants responded explicitly by pressing one of two keyboard buttons marked "yes" or "no." These explicit ratings of document relevance by participants are called *perceived relevance* in this article and are the focus of our analysis. In presenting the results, we call documents judged as relevant *Rp-trials* and documents judged irrelevant *Ip-trials*. We do not report data from the covert blocks in this article.

Document Corpus

Questions and relevance assessments were taken from the TREC 2005 Q&A track (Voorhees, 1998) which used AQUAINT corpus news stories (Graff, 2002). We selected 65 news stories (63 IS task, two training) with low text length variation and one or two paragraphs. The TREC relevance assessments were manually verified. The word screen coordinates were automatically extracted to match the eye fixations on the relevant words.

Data Preprocessing & Analysis

Data Cleanup

After data cleanup, our analysis was performed on 744 trials. Table 1 summarizes the steps.

TABLE 1. Data cleanup summary.

Data cleanup step #	Trials	Quantity
0. Original data	Total number of trials	1,008 (24 participants * 42 overt trials)
1. EYE data cleanup	Trials with good EYE data	907
2. removal of irrelevant docs	Trials balanced with respect to doc. relevance	791
3. EEG data missing for 1 person	Total trials with recorded EEG data	945 (no data for 1 participant)
4. EEG data cleanup	Trials with good EEG data	911
5. after EYE and EEG cleanup	Total trials (EYE + EEG) with good data	744

Step 1: Poor-quality EYE data records (Tobii validity <4) and off-screen fixations were removed (~5% of fixations). Minimal continuity of eye-tracking data was enforced. We removed trials with gaps longer than 1 second (Oliveira et al., 2009), or >20% missing data or fewer than four fixations on a document. Remaining trials with small EYE data gaps (possibly a result of eye blinks) were treated by merging pupil measurements for both eyes, using the mean if both were captured, or by using one eye if the other eye was "lost." A linear interpolation algorithm was applied to the remaining pupil data gaps.

Step 2: The relevance level (I/T/R) distribution of the remaining trials was not equal. To gain a more uniform distribution of document relevance degrees and text lengths, we removed data from "long" (≥310 words) irrelevant documents to better match the document characteristics of the remaining topical and relevant trials. The resulting trial collection had an average document length of 178 words (SD = 30) with relevance level distribution: I/T/R (19/21/19).

EEG data issues affected 34 trials (Step 4, below). It is reasonable to think the missing observations were random because the dropped trials were a result of data collection errors.

The following preprocessing steps were performed on EEG data:

- Data for channels 5, 10 (T7, T8 Figure 1) was re-referenced.
- Applied a 1–40 Hz bandpass filter.
- Signal mean was removed (trimmed mean for middle 90% percentile).
- ICA analysis was performed and artifacts (eye movement, blinks, etc.) removed.

Artifact Removal

Raw EEG data contains artifacts from eye movements, facial muscle contractions, and electric device interference. Independent component analysis (ICA) was used to decompose brain signals into signal components associated with the artifact generators using pattern and artifact source localization analysis. Some artifact sources (e.g., eye blinking, muscle movements) present challenges for across-participants analysis because they are user dependent. We used ADJUST (Mognon, Jovicich, Bruzzone, & Buiatti, 2011) to process our data.

EYE Features

Eye movements were analyzed using a model (Cole et al., 2011; Cole, Gwizdka, Liu, Belkin, & Zhang, 2013) influenced by E-Z Reader (Reichle, Rayner, & Pollatsek, 2003). The reading model assumes words are processed serially one at a time. A single fixation can result in processing more than one word when the next word in the reading direction is identified in parafoveal view (Rayner, 1975; Schotter, Angele, & Rayner, 2011). The model assumes a minimum fixation time (150 ms) is required to acquire word meaning. These assumptions have strong empirical support. Fixations of >150 ms were grouped into continuous reading sequences (labeled "R") or isolated scanning fixations ("S"). Saccade distances were calculated for reading sequences.

Under constant illumination, pupil dilation is associated with several cognitive functions, including interest (Krugman, 1964) and changes in attention (Hoeks & Levelt, 1993; Wierda, Rijn, Taatgen, & Martens, 2012). We controlled lab environment lighting and text document formatting (black background, white font, and low word number variability) to achieve almost no variability of luminance across all trials. We removed individual variability in pupil sizes and pupillary response. We calculated relative change in pupil dilation (RPD^i_t) from pupil measurement at a time t P_t and participant pupil baseline ($P^i_{baseline}$), where baseline is the average pupil size over all text document presentations (Eq. 1).

$$RPD^i_t = (P_t - P^i_{baseline}) / P^i_{baseline} \quad (1)$$

We calculated 25 eye-tracking data features (Table 2).

EEG Features

EEG signal features were extracted to capture brain activity (Table 3).

To represent an EEG data epoch, univariate features were extracted from each EEG channel and concatenated with the multivariate features to make a feature vector. EEG signals have high individual variability, so for across-participants analysis we performed personalized feature standardization (Wang et al., 2015).

Data Segmentation for Classification

Intuition and initial data exploration suggested that initial document reading would be similar across the relevance conditions. Changes in reading relevant and irrelevant documents, and the associated cognitive processes, were expected to occur both later and closer to the relevance decision. We expected variation for relevance decision timing because the location of the relevant

TABLE 2. Eye-tracking features (25 features in total).

Group	Feature	Description
Pupil	RPD	Relative change in pupil diameter, normalized to a user's mean value: Mean, STD, and 5 percentiles (10 25 50 75 90)
	RPD moving-average	Moving average of the 1st order difference (gradient) of RPD with a lag of 10: Mean, SD, min, max, and 3 percentiles (25 50 75)
Reading sequence	Fixation duration	Mean, SD, Sum of fixation durations
	Fixation count	Number of fixations in a reading sequence
	Saccade distance	Mean, SD, Sum of Euclidian distances between fixations in a sequence
Scanning fixations	Fixation duration	Mean, SD, sum of fixation durations
	Fixation count	Total number of isolated scanning fixations

TABLE 3. EEG features (50 types of features in 16 groups; 569 features in total).

Feature	Description
Signal statistics	Mean, SD, skewness, kurtosis
Morphological features	Curve length, zero crossings, number of peaks
Average nonlinear energy	A measure sensitive to signal power spectral changes
Power spectral features	50%, 90%, 95% spectral edge frequency
Band power	For four frequency bands*
Relative band power	Power ratios for each frequency band*
Interhemispheric asymmetry	Power ratios between right-left hemisphere channel groups** for 4 freq. bands*
Intrahemispheric asymmetry	Power ratios within right-left hemisphere channel groups** for 4 freq. bands* (Cvetkovic & Cosic, 2009)
Hurst exponent	Measures EEG nonstationary dynamics (Hwa & Ferree, 2002)
Hjorth measures	Activity, Mobility, Complexity (Hjorth, 1970)
Barlow measures	Mean amplitude, mean frequency, and spectral purity index (Goncharova & Barlow, 1990)
Wavelet entropy	Calculated from wavelet coefficients; indicates the degree of multi-frequency signal order/disorder in EEG.
Relative wavelet energy	For four frequency bands*
Ratio of Alpha to Beta Band Power	Relative changes between alpha and beta band signals
Range to Variance Ratio	
Connectivity network features	Fraction of channel pairs with correlations $>$ C which can be 0.1, 0.2, . . ., and 0.9. The ratio indicates overall EEG connectivity strength given a value of C.

*theta: 4–8 Hz, alpha: 8–13 Hz, beta: 13–25 Hz, low gamma: 25–40 Hz.

**Four channel groups: frontal-left (AF3, F7, F3, FC5), frontal-right (FC6, F4, F8, AF4), back-left (P7, O1), and back-right (P8, O2).

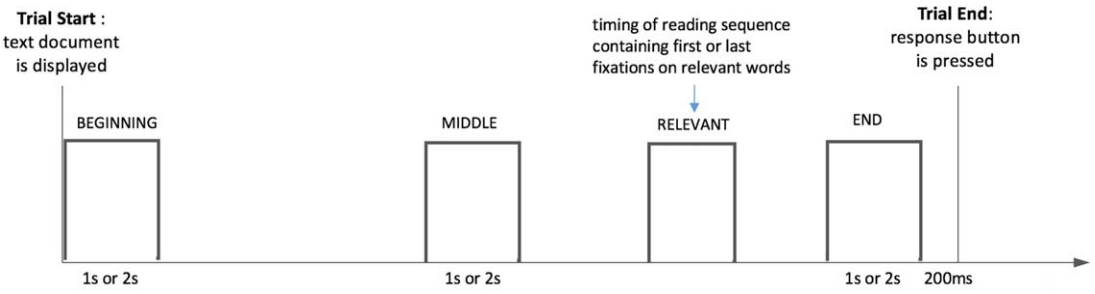

FIG. 3. Illustration of the data epochs with respect to timing of a trial. [Color figure can be viewed at wileyonlinelibrary.com]

information (if any) was not controlled. Given the large variation in trial reading time, we selected from each trial three representative equal-length time epochs positioned at the beginning, in the middle, and at the end of a trial (Figure 3). Selecting data epochs from such different parts of trials enables us to investigate temporal dynamics of eye-tracking and EEG signals. To address RQ1, we compared binary classification model performance (for perceived relevant vs. irrelevant text) learned in each data epoch (*classifications A*). Additional epochs were defined based on the timing of reading sequences that contained either the first (first-Rfix) or the last (last-Rfix) eye-fixations on relevant words (in Rp-trials). Data from each such segment is compared, separately, with data from beginning, middle and end of Ip-trials. Notice that a reading sequence represents plausibly a coherent unit of reading more so than isolated fixations on individual words, so epoch analysis using reading sequences is interesting. These data epochs were used in classifications denoted "*B*" to investigate RQ2. The epochs are illustrated in Figure 3.

We investigated epoch classification performance changes using two epoch lengths, 1,000 ms and 2,000 ms,

TABLE 4. Classifications A (epoch 2,000 ms-long).

Feature	Compared Epochs	AUC	ACC	Sen+Spe 2	Sensitivity	Specificity
EEG	Beg	0.57	0.55	0.53	0.58	0.47
	Mid	0.60	0.60	0.58	0.62	0.55
	End	0.59	0.65	0.54	0.83	0.24
EYE	Beg	0.56	0.40	0.52	0.19	0.85
	Mid	0.70	0.66	0.65	0.67	0.63
	End	0.80	0.72	0.71	0.74	0.67
EEG+EYE	Beg	0.55	0.46	0.52	0.35	0.69
	Mid	0.70	0.66	0.65	0.69	0.60
	End	0.79	0.71	0.71	0.73	0.68

TABLE 5. Classifications B.

Features	Ip-trial epoch	AUC	ACC	(Sen+Spe)/2	Sensitivity	Specificity	# Rp-trials	# Ip-trials
				Epoch 1,000 ms-long.				
EEG	*beg*	0.56	0.57	0.53	0.66	0.41	165	290
	end	0.64	0.60	0.59	0.61	0.57	165	279
EYE	*beg*	**0.95**	**0.86**	**0.87**	**0.84**	**0.89**	165	290
	end	0.78	0.72	0.72	0.74	0.69	165	279
EEG+EYE	*beg*	**0.96**	**0.87**	**0.88**	**0.86**	**0.90**	165	290
	end	0.78	0.72	0.71	0.73	0.70	165	279
				Epoch 2,000 ms-long.				
EEG	*beg*	**0.79**	**0.81**	**0.69**	**0.93**	**0.45**	165	290
	end	0.77	0.76	0.69	0.97	0.41	165	279
EYE	*beg*	**0.88**	**0.79**	**0.79**	**0.81**	**0.76**	165	290
	end	0.77	0.71	0.71	0.74	0.68	165	279
EEG+EYE	*beg*	**0.91**	**0.83**	**0.82**	**0.89**	**0.75**	165	290
	end	0.85	0.80	0.76	0.83	0.69	165	279

which are expected to reflect signal dynamics during relevance decisions. Prior EEG research showed relevant stimulus discrimination at 300–800 ms after onset (Allegretti et al., 2015; Eugster et al., 2014; Frey et al., 2013). Eye-tracking research has shown the last 1,000 ms and 2,000 ms in a trial as significant (Gwizdka, 2014; Gwizdka & Zhang, 2015). Data sampling rates were 128 Hz (EEG) and 60 Hz (eye-tracker). The mean fixation duration was 239 ms ($SD = 91$ ms). Given the expected rise time of EEG relevance signals (<1,000 ms), we set 1,000 ms as the shortest epoch length. End-trial epochs avoided motor control brain activity (key press) by removing the last 200 ms.

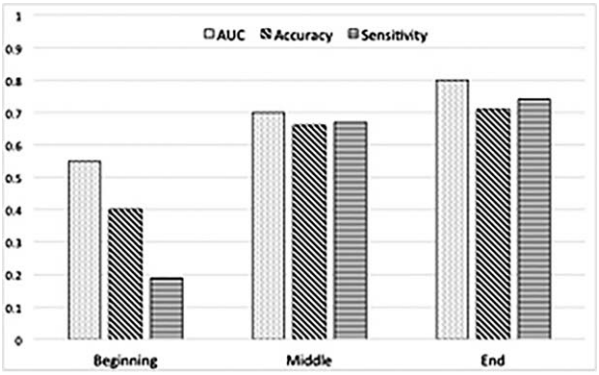

FIG. 4. Conceptual illustration of diverging differences in Classifications A results.

Feature Selection

We used 569 EEG features (Table 3) and 25 EYE features (Table 2). Minimum redundancy maximum relevance (mRMR; Peng, Long, & Ding, 2005) was used to identify the most informative features associated with relevance decisions. mRMR selects the most relevant feature subset with minimal redundant features using mutual information analysis and is frequently used in bioinformatics (Saeys, Inza, & Larrañaga, 2007).

Classification Method

Proximal Support Vector Machines (PSVM; Fung & Mangasarian, 2001) were used to model EEG and EYE feature patterns for relevant and irrelevant documents. Three feature sets were used (EEG, EYE, and EEG + EYE features). Binary classification models (target: perceived relevance) were created for epoch and trial selection combinations. Input was from data epochs using the 12 different timing criteria (see Data Segmentation for Classification). Models were constructed from all trials, and trials with correct responses. We constructed 162 perceived relevance classification models (3 epochs * 12 timing criteria * 3 feature sets * 2 trial selection criteria) from the EEG and eye-tracking signals.

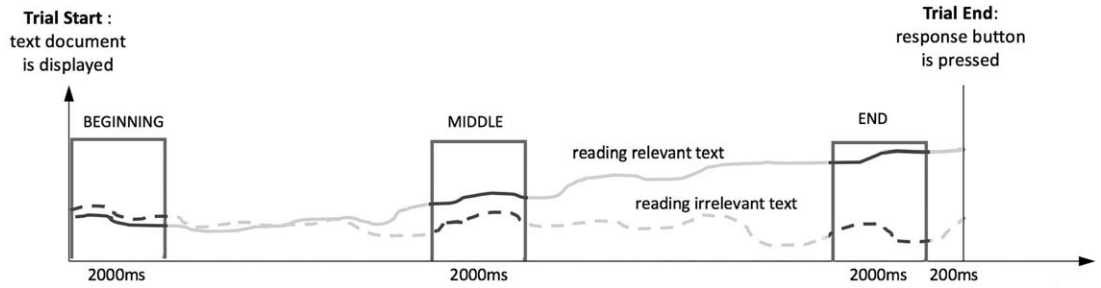

FIG. 5. Comparison of classification A performance for beginning, middle, and end epochs of eye data.

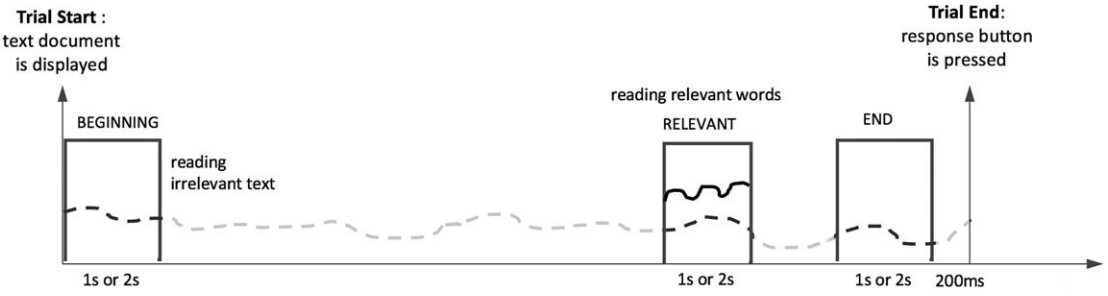

FIG. 6. Conceptual illustration of Classifications B.

Training and Evaluation

The PSVM classifiers were trained and evaluated using 10-fold cross-validation. Model sensitivity and specificity (Eqs. 2, 3) were used to evaluate the classification performance. Overall performance was measured as the average of sensitivity and specificity.

$$sensitivity = \frac{\#\ of\ correctly\ classified\ samples\ of\ relevent\ judgement}{total\ number\ of\ samples\ of\ relevent\ judgement} \quad (2)$$

$$specificity = \frac{\#\ of\ correctly\ classified\ samples\ of\ irrelevent\ judgement}{total\ number\ of\ samples\ of\ irrelevent\ judgement} \quad (3)$$

Results

Participant RJ accuracy was 90.3%, confirming the tasks were performed as expected. The average text reading time was 10.4 sec ($SD = 5.2$ sec), well within the 20-second limit, and is evidence there was little time stress on participant RJ.

The binary classification of the perceived relevance of the whole trial EEG and EYE data was: EEG features AUC = 0.60; accuracy (ACC) = 0.54–0.60; for EYE features AUC = 0.76–0.78 and ACC = 0.69–0.71. EEG-only classification was barely better than random, whereas the EYE classification was reasonably good. This is evidence that relevance judging effects may be detectable in EYE data but not EEG signals at the level of reading whole text documents.

Tables 4 and 5 present epoch-level classification model performance of two types for the trials with correct responses. Classifications A (Table 4) compares Rp-trials with Ip-trials separately within each of the three epochs: beginning, middle, and end. Classification performance tends to improve from beginning to final epoch (Figure 4), and is strongest for EYE (final epoch AUC = 0.80, ACC = 0.72). Figure 5 shows comparison of classification A performance for EYE data.

TABLE 6. Three best EYE features for Classifications A (epoch 1,000 ms-long).

	p values for epochs		
Best features	beg	middle	end
RPD – 25th percentile	0.54	$6*10^{-6}$ **	$1.3*10^{-11}$ ***
RPD – mean	0.50	$2*10^{-5}$ **	$2.1*10^{-12}$ ***
RPD – 10th percentile	0.47	$3*10^{-5}$ **	$8*10^{-9}$ ***

Significant differences at: $**p < .001$; $***p < 10^{-8}$.

Classifications B compared first-Rfix and, separately, last-Rfix epochs with beginning and end data epochs on Ip-trials (Table 5). Classification performance is better for last fixations than for first fixations on relevant words and therefore we present only data for last-Rfix epochs. Classification B epochs are illustrated in Figure 6.

Table 5 shows that EYE feature classification models performed better for 1,000 ms-long epochs, whereas classification of EEG performed better for 2,000 ms-long epochs. This may indicate that EYE and EEG models "tune" to different cognitive processes or phases of a process. Classification performance is better in comparing relevant epochs in Rp-trials with beginning epochs vs. final epochs in Ip-trials.

TABLE 7. *t*-tests for best six features (EEG, EYE).

EEG	Feature	ZC,Ch:P8	WE,Ch:FC5	Kurtosis,Ch:FC5	ZC,Ch:T7	WE,Ch:P7	ZC,Ch:O2
	p-value	$1.17*10^{-24}$	$7.8*10^{-7}$	$1.48*10^{-6}$	$6.16*10^{-26}$	$5.3*10^{-8}$	$5.52*10^{-25}$
EYE	Feature	Reading-total duration	Reading-total distance	Scanning fixation duration *SD*	Reading fixation duration *SD*	Scanning total duration	Reading-distance *SD*
	p value	$1.23*10^{-13}$	$3.68*10^{-15}$	$1.55*10^{-15}$	$2.89*10^{-15}$	$5.15*10^{-18}$	$1.92*10^{-10}$

Note. *SD* = standard deviation.

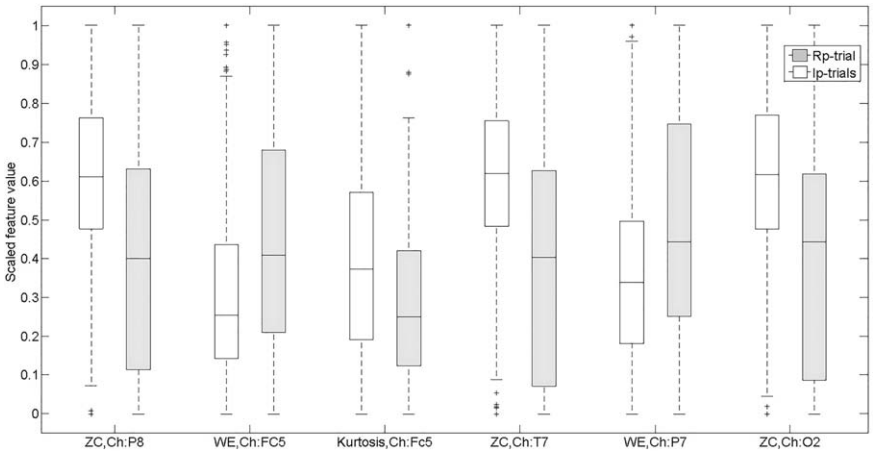

FIG. 7. Boxplots for the top six EEG features for Rp-trials (last Rfix) versus Ip-trials (ending epoch): Classifications B.

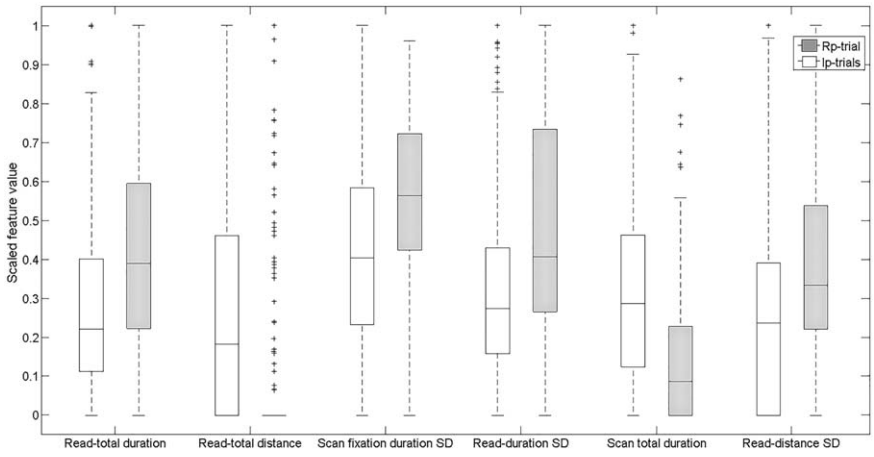

FIG. 8. Boxplots for the top six EYE features for Rp-trials (last Rfix) versus Ip-trials (ending epoch): Classifications B.

Adding EEG features to EYE features does not improve models at 1,000 ms-long epochs, but mildly improves 2,000 ms-long epoch models (0.03–0.08).

Best Features

Important model features were calculated for selected best perceived relevance prediction models (see Feature Selection). The better performing classifications A were for EYE features. The *t*-tests show significant differences between the best feature values for Rp-trials and Ip-trials for middle and end epochs, but not for beginning epochs (Table 6).

The best performing classifications B were for EEG + EYE for last-Rfix epochs vs. reading beginning of Ip-trials (lower-part of Table 5). For these epochs, the six most important EYE and EEG features were calculated (Table 7). Differences between such features were statistically significant (Table 7) and are shown for normalized feature values in Figures 7 and 8.

The best EEG features were kurtosis and wavelet entropy (WE) for channel FC5 and zero-crossing (ZC) for channel

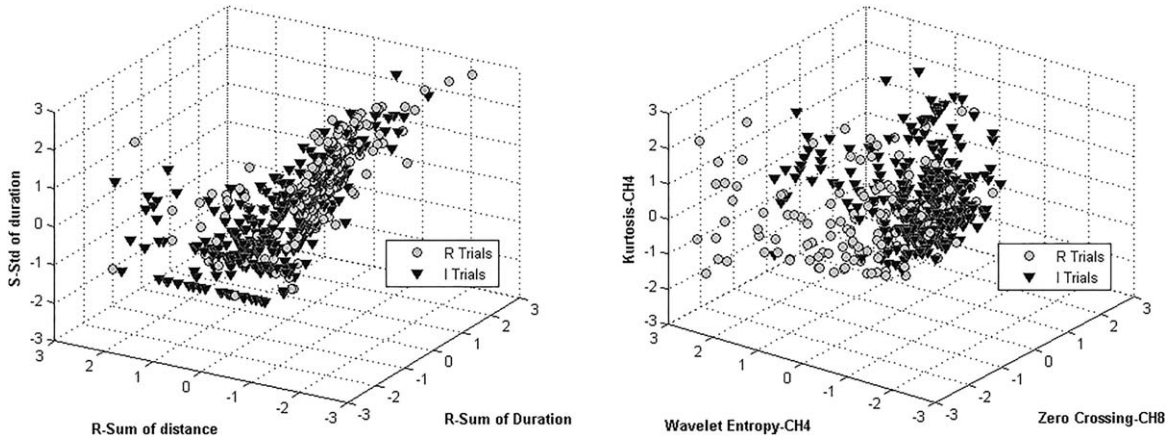

FIG. 9. Three EYE (left) and EEG (right) features that best discriminate between last-Rfix and the reading at the end on an Ip-trial.

P8. The WE characterizes the order or disorder of signal power for the five brain signal frequency bands (delta, theta, alpha, beta, gamma). We may be sensitive to capturing brain dynamic changes in the text reading task (e.g., related to changes in mental load). Frequency of zero-crossing tends to change with changes between mental states. High kurtosis has more peaks as compared to the baseline. The kurtosis in Rp-trials was higher than in the Ip-trials. The best EYE features were related to reading type (i.e., continuous reading vs. scanning). Figure 9 shows, separately for EYE and EEG, how the three best features discriminate between Rp-trials and Ip-trials.

Discussion

We started with three research questions. RQ1 concerned differences in the dynamics of eye-tracking and EEG signals in reading relevant and irrelevant documents. RQ2 concerned differences between eye-tracking and EEG signals during reading relevant words vs. reading irrelevant documents. Finally, in RQ3 we wanted to learn the potential of a low-cost wireless EEG device to predict perceived relevance. We focus on perceived relevance because of the association with the participant's cognitive processes during the trials.

For RQ1, we found interesting results in the temporal dynamics of eye-tracking and EEG signals at the three different temporal stages of relevant and irrelevant document processing (Classifications A). As reading progressed there was an increasing difference in performance of classification A for beginning, middle, and epochs. This difference was particularly evident in EYE-feature classification models, and was also seen in EEG-feature models. Differences in pupil dilation and fixation-based measures between relevant and relevant documents increased as the documents were read (Figure 4). These may reflect differences in cognitive processes detected by eye-tracking and EEG.

For RQ2 and classifications B, the EYE and EEG models show differences in reading relevant words compared to reading parts of irrelevant text. EYE models provided good classification performance for perceived relevance (best for 1,000 ms, AUC = 0.95 and ACC = 0.86, upper part of Table 5). This is evidence that reading relevant text affects eye movement patterns and pupil dynamics. The greatest difference for eye-tracking data was found between the last pass through relevant words in Rp-trials (last-Rfix) and the beginning of Ip-trials. Differences between reading relevant words and the ending epochs of Ip-trials were still large (AUC = 0.77–0.78).

The Classifications B results provide evidence that EEG signals can distinguish RJs. The best performance levels (AUC = 0.79 and AUC = 0.81) were found for final reading of relevant words (last-Rfix) versus reading beginning of an irrelevant text for 2,000 ms-long epochs (lower part of Table 5). Classification performance was diminished for other epochs. One possible explanation is an RJ was made during a final pass through relevant words, resulting in detection of cognitive activity that differed from that of reading text.

Does the addition of EEG data improve the EYE models? The best combined EYE + EEG models (best AUC = 0.96 and ACC = 0.87; for 1,000 ms epochs) were derived for the cases that also had the best EYE models. The performance improvement was only 0.01–0.02 (for AUC/ACC). However, in other cases combined EYE + EEG data increased classification performance by 0.01–0.08 as compared to EYE data alone. These were cases where EYE feature classification performance was lower and more similar to EEG performance (e.g., for 2,000 ms-long epochs and for end epoch for Ip-trials with epoch from last fixation on relevant words with AUC = 0.77–0.78 for EYE, AUC = 0.77–0.79 for EEG; lower part of Table 5). Thus, there is a positive result for RQ3 but only for selected situations.

Relating the EEG model and feature results to specific brain areas and their function is not well supported by our method and the Emotiv EEG device. Scalp electrodes do not reflect particular areas of cortex, as the active sources are hard to localize due to nonhomogeneous skull properties and orientation of the cortex sources (Nunez, 2002). Despite these limitations, we can speculate and relate EEG electrode location to basic brain function. We found the most

discriminative features were from channels P7, P8, O2, T7, and FC5, that is, from frontal-central (FC), temporal (T), parietal (P), and occipital (O) lobes (Figure 1). Based on localization of Brodmann areas (2007), one can speculate that brain activity captured by channel FC5 may be related to verbal reasoning; T7 to certain memory functions and auditory processing; T7 and P7 to verbal and reading comprehension (near Wernicke's area; DeWitt & Rauschecker, 2013); and O2 to visual processing.

The 1- to 2-second epoch classification results were superior to the whole trial results for both devices. The shorter epochs match better with the timespan of distinct cognitive processes. Data collected over longer periods likely mixes the activity of multiple processes. This points us towards ERP study design and data analysis, but a low-cost EEG device may be too limited (Duvinage et al., 2013). Sliding time window analysis may be an effective procedure to infer relevance in near real time and this is one plan for future work. We also plan to study longer text documents, more complex search tasks, and multistage search sessions.

Taken together, the results provide an insight into differences between eye-tracking and EEG, and, possibly, into the differences between cognitive processes involved in reading text documents and judging their relevance. The eye-tracking data provides good support in this regard. The EEG results with a low-cost wireless device were consistent with EYE data. It seems fair to say that this type of EEG device can detect differences in judging relevant and irrelevant documents when distinctly different cognitive processing is used, as in relevance decision-making.

The best EYE features overlap with those reported in Gwizdka (2014) for classifications at the level of whole document trials. The two sets of the selected best EYE features for Classifications A and B are different (Tables 6 and 7). They reflect characteristics of pupil dilation vs. reading patterns (i.e., continuous reading and scanning), respectively. Pupil dilation increased as participants continued to read relevant documents. Reading relevant words differed from reading irrelevant text in longer fixations and higher propensity for reading than for scanning (Figure 8).

The differences in the best models by epoch is an especially encouraging result of this work because of real-world noise that might be expected to make detection of cognitive processing of relevant documents difficult. For relevant documents the location of the relevant words in one document might be near the beginning of the text and in another in the middle or near the end. This reflects real-world information search. Likewise, one expects some variance in user behaviors after processing relevant words. For example, one might be biased to stop reading immediately, whereas another might continue to read for a while. All of these real-world situations introduce noise into our modeling of the processing of relevant and irrelevant documents.

Despite this noise, we obtained models that can discriminate between the time series observations of reading relevant vs. irrelevant documents with expected divergences in classification performance. Model performance should be improved by taking account of individual differences in reading patterns of relevant and irrelevant documents. Improved models might also further distinguish the shift in feature importance. The results provide clear evidence that participants were performing some cognitive process in a distinguishably different way in the late epoch depending on the relevance of the document. As shown in Figures 7 and 8, the differences in values of the top six EEG and EYE features extracted from epochs starting at the onset of reading last relevant words vs. ending epochs on Ip-trials are statistically significant.

The divergence in measurements (EEG + EYE) when relevant text was being processed as compared to processing irrelevant text (i.e., when no positive RJ was made) indicates the user might be in a distinguishably different state while reading relevant text and, in particular, before and after the RJ. These changes of state are reflected in changed reading behaviors and brain activity. Such brain activity and behavior differences might be explained by changes in the user's knowledge and changes in immediate or overall task goals. Our study was not able to address these speculations because the experiment blocks (each with a question and three texts that might contain an answer) were fixed. A future study could be designed as a multistage task where identification of a relevant answer to the initial question would invoke a knowledge gain that enables understanding of the task requirements for a follow-up question.

A challenge is to identify user responses when they encounter the relevant information. One could be user's satisfaction with achieving a task goal; another, learning and gaining a new perspective on the task goal because of gained knowledge. Plausibly, the first has an affective component, whereas the second cognitive. NP research tools provide ways to investigate differences between affective causes of search behaviors and cognitive effects. Other data collection methods, such as self-reporting, think-aloud protocols, or posttask reflection are subjective. This indicates the potential for NP-methods to contribute to the foundations of IS&R. We term this new approach *neuro-information science*.

Limitations

Tobii T60 eye-tracker has a reported accuracy of 0.5°; in practice, the accuracy may be lower. It limited our ability to identify fixations on relevant words in all trials (Rp-trials success: 72%). Classifications were performed only across all participant data. The potentially significant negative effects of individual differences were mitigated by performing personalized feature standardization.

This study focused entirely on topical RJs and used tasks similar to the Cranfield single task query and document unit. Such judgments are an important element of many search sessions. We believe that this narrower focus of our study is a good starting point for further research that can investigate the richness of cognitive processes in extended complex search sessions.

Conclusions

To our knowledge, there is no prior work applying this kind of relevance analysis to EEG and eye-tracking data. This is our main IS&R contribution. Our work is exploratory in nature, and we have offered speculative explanations for the EEG and EYE perceived relevance classification results. This is a view of temporal reading process dynamics in the context of a relevance decision process. The selected eye-tracking features seem sensitive to both processes, whereas the EEG features derived from our EEG device work mainly for the relevance decision process.

Examination of the selected best predictor features are cursory. This is a direction for future work. Also, the Results and Discussion sections were devoted exclusively to perceived relevance. Our extended analysis, not reported here, showed similar classification performance is possible for document relevance. We believe perceived relevance is more important for understanding cognitive processes at the foundation of RJs. As a general point, to gain a deeper understanding of relevance and its associated cognitive processes it is best to focus on user-subjective processes and their observable measures. NP methods, such as those applied in our work, offer the prospect of looking "into the user's brain." Progress in NP methods and results from cognitive neuroscience can enrich IS&R research programs in transformative ways through understanding the cognitive foundations that shape user interactions with information.

This work contributes more generally to the IS&R field by developing an objective method to investigate processes of reading and subjective RJs in search. It addresses IS&R foundations by exploring the empirical-grounding for a key IS&R construct, relevance, in a neural-centered description of cognitive processes. There is potential for practical application of such techniques. For example, near real-time detection of RJs within search sessions may be achievable. This could drive system personalization, search intent detection, and relevance feedback to search algorithms.

Acknowledgments

This work was supported, in part, by Google-Faculty-Research-Award to Jacek Gwizdka.

References

Abbott, W.W., & Faisal, A.A. (2012). Ultra-low-cost 3D gaze estimation: An intuitive high information throughput compliment to direct brain–machine interfaces. Journal of Neural Engineering, 9, 046016.

Ajanki, A., Hardoon, D., Kaski, S., Puolamäki, K., & Shawe-Taylor, J. (2009). Can eyes reveal interest? Implicit queries from gaze patterns. User Modeling and User-Adapted Interaction, 19, 307–339.

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 SIGIR'2015 (pp. 719–722). New York: ACM.

Anderson, E.W., Potter, K.C., Matzen, L.E., Shepherd, J.F., Preston, G.A., & Silva, C.T. (2011). A user study of visualization effectiveness using EEG and cognitive load. Computer Graphics Forum, 30, 791–800.

Back, J., & Oppenheim, C. (2001). A model of cognitive load for IR: Implications for user relevance feedback interaction. Information Research, 6.

Badcock, N.A., Mousikou, P., Mahajan, Y., de Lissa, P., Thie, J., & McArthur, G. (2013). Validation of the Emotiv EPOC EEG gaming system for measuring research quality auditory ERPs. PeerJ, 1, e38.

Balatsoukas, P., & Ruthven, I. (2012). An eye-tracking approach to the analysis of relevance judgments on the Web: The case of Google search engine. Journal of the Association for Information Science and Technology, 63, 1728–1746.

Barry, C.L. (1994). User-defined relevance criteria: An exploratory study. Journal of the American Society for Information Science and Technology, 45, 149–159.

Bobrov, P., Frolov, A., Cantor, C., Bakhnyan, M., & Zhavoronkov, A. (2011). Brain-computer interface based on generation of visual images. PLoS One, 6, e20674.

Borlund, P. (2003). The concept of relevance in IR. Journal of the American Society for Information Science and Technology, 54, 913–925.

Brodmann, K. (2007). Brodmann's: Localisation in the cerebral cortex. New York: Springer.

Brouwer, A.-M., Reuderink, B., Vincent, J., van Gerven, M. A., & van Erp, J. B. (2013). Distinguishing between target and nontarget fixations in a visual search task using fixation-related potentials. Journal of Vision, 13, 17.

Buscher, G., Dengel, A., Biedert, R., & Elst, L.V. (2012). Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond. ACM Transactions on Interactive Intelligent Systems, 1, 9:1–9:30.

Cole, M.J., Gwizdka, J., Liu, C., Belkin, N.J., & Zhang, X. (2013). Inferring user knowledge level from eye movement patterns. Information Processing and Management, 49, 1075–1091.

Cole, M.J., Gwizdka, J., Liu, C., Bierig, R., Belkin, N.J., & Zhang, X. (2011). Task and user effects on reading patterns in information search. Interacting with Computers, 23, 346–362.

Cooke, L. (2006). Is the mouse a "poor man's eye tracker"? In Proceedings STC'2006 (Vol. 53, p. 252).

Cvetkovic, D., & Cosic, I. (2009). EEG inter/intra-hemispheric coherence and asymmetric responses to visual stimulations. Medical & Biological Engineering & Computing, 47, 1023–1034.

DeWitt, I., & Rauschecker, J.P. (2013). Wernicke's area revisited: Parallel streams and word processing. Brain Language, 127, 181–191.

Dupret, G., & Liao, C. (2010). A model to estimate intrinsic document relevance from the Clickthrough logs of a web search engine. In Proceedings of WSDM'2010 (pp. 181–190). New York: ACM.

Duvinage, M., Castermans, T., Petieau, M., Hoellinger, T., Cheron, G., & Dutoit, T. (2013). Performance of the Emotiv EPOC headset for P300-based applications. BioMedical Engineering OnLine, 12, 1–15.

Eugster, M.J.A., Ruotsalo, T., Spapé, M.M., Kosunen, I., Barral, O., Ravaja, N., ... Kaski, S. (2014). Predicting term-relevance from brain signals. In Proceedings of SIGIR'2014 (pp. 425–434). New York: ACM.

Fitzgerald, M.A., & Galloway, C. (2001). Relevance judging, evaluation, and decision making in virtual libraries: A descriptive study. Journal of the American Society for Information Science and Technology, 52, 989–1010.

Frey, A., Ionescu, G., Lemaire, B., López-Orozco, F., Baccino, T., & Guérin-Dugué, A. (2013). Decision-making in information seeking on texts: An eye-fixation-related potentials investigation. Frontiers in Systems Neuroscience, 7, 39.

Fung, G., & Mangasarian, O.L. (2001). Proximal support vector machine classifiers. In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 77–86). New York: ACM.

Goncharova, I.I., & Barlow, J.S. (1990). Changes in EEG mean frequency and spectral purity during spontaneous alpha blocking. Electroencephalography and Clinical Neurophysiology, 76, 197–204.

Graff, D. (2002). The AQUAINT corpus of English news text. Retrieved from http://www.language-archives.org/item/oai:www.ldc.upenn.edu:LDC2002T31

Guo, Q., & Agichtein, E. (2010). Towards predicting web searcher gaze position from mouse movements. In Proceedings CHI'2010 Extended Abstracts on Human Factors in Computing Systems (pp. 3601–3606). New York: ACM.

Gwizdka, J. (2014). Characterizing relevance with eye-tracking measures. In Proceedings of IIiX'2014 (pp. 58–67). New York: ACM.

Gwizdka, J., & Zhang, Y. (2015). Differences in eye-tracking measures between visits and revisits to relevant and irrelevant web pages. In Proceedings of SIGIR'2015 (pp. 811–814). New York: ACM.

Healy, G., & Smeaton, A.F. (2011). Eye fixation related potentials in a target search task. In Proceedings of IEEE EMBC'2011 (pp. 4203–4206).

Hjørland, B. (2010). The foundation of the concept of relevance. Journal of the American Society for Information Science and Technology, 61, 217–237.

Hjorth, B. (1970). EEG analysis based on time domain properties. Electroencephalography and Clinical Neurophysiology, 29, 306–310.

Hoeks, B., & Levelt, W.J.M. (1993). Pupillary dilation as a measure of attention: A quantitative system analysis. Behavior Research Methods, Instruments, & Computers, 25, 16–26.

Huang, J., White, R., & Buscher, G. (2012). User see, user point: Gaze and cursor alignment in web search. In Proceedings of CHI'2012 (pp. 1341–1350). New York: ACM.

Huang, X., & Soergel, D. (2013). Relevance: An improved framework for explicating the notion. Journal of the American Society for Information Science and Technology, 64, 18–35.

Hwa, R.C., & Ferree, T.C. (2002). Scaling properties of fluctuations in the human electroencephalogram. Physical Review E, 66, 021901.

Jung, S., Herlocker, J.L., & Webster, J. (2007). Click data as implicit relevance feedback in web search. Information Processing & Management, 43, 791–807.

Just, M.A., & Carpenter, P.A. (1987). The psychology of reading and language comprehension. Needham Heights, MA: Allyn & Bacon.

Kellar, M. (2004). Effect of task on time spent reading as an implicit measure of interest. In Proceedings of ASIS&T'2004 (Vol. 41, pp. 168–175).

Kelly, D., & Belkin, N.J. (2001). Reading time, scrolling and interaction: Exploring implicit sources of user preferences for relevance feedback. In SIGIR Forum (pp. 408–409).

Kelly, D., & Belkin, N.J. (2004). Display time as implicit feedback: Understanding task effects. In Proceedings of SIGIR'2004 (pp. 377–384). New York: ACM.

Khushaba, R.N., Greenacre, L., Kodagoda, S., Louviere, J., Burke, S., & Dissanayake, G. (2012). Choice modeling and the brain: A study on the Electroencephalogram (EEG) of preferences. Expert Systems with Applications, 39, 12378–12388.

Kim, M., Kim, B.H., & Jo, S. (2015). Quantitative evaluation of a low-cost noninvasive hybrid interface based on EEG and eye movement. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 23, 159–168.

Knoll, A., Wang, Y., Chen, F., Xu, J., Ruiz, N., Epps, J., & Zarjam, P. (2011). Measuring cognitive workload with low-cost electroencephalograph. In P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, & M. Winckler (Eds.), Proceedings of INTERACT 2011 (Vol. 6949, pp. 568–571). Berlin: Springer.

Krugman, H.E. (1964). Some applications of pupil measurement. Journal of Marketing Research, 1, 15.

Lesk, M.E., & Salton, G. (1968). Relevance assessments and retrieval system evaluation. Information Storage and Retrieval, 4, 343–359.

Liu, C., Liu, J., & Belkin, N.J. (2014). Predicting search task difficulty at different search stages. In Proceedings of CIKM'2014 (pp. 569–578). New York: ACM.

Liu, J., & Belkin, N.J. (2010). Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. In Proceedings of SIGIR'2010 (pp. 26–33). New York: ACM.

Marcos, M.-C., Gavin, F., & Arapakis, I. (2015). Effect of snippets on user experience in web search. In Proceedings of the XVI International Conference on HCI (pp. 47:1–47:8). New York: ACM.

Mognon, A., Jovicich, J., Bruzzone, L., & Buiatti, M. (2011). ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. Psychophysiology, 48, 229–240.

Moshfeghi, Y., Pinto, L.R., Pollick, F.E., & Jose, J.M., (2013). Understanding relevance: An fMRI study. In P. Serdyukov, P. Braslavski, S.O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, & E. Yimaz (Eds.), Advances in information retrieval (pp. 14–25). Berlin: Springer.

Nunez, P.L. (2002). Electroencephalography (EEG). In Encyclopedia of the human brain (pp. 169–179). New York: Academic Press.

Oliveira, F.T.P., Aula, A., & Russell, D.M. (2009). Discriminating the relevance of web search results with measures of pupil size. In Proceedings CHI'2009 (pp. 2209–2212). Boston: ACM.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis, 27, 1226–1238.

Park, H., & Reder, L. (2004). Moses illusion: Implication for human cognition. In R.H. Pohl (Ed.), Cognitive illusions (pp. 275–291). Hove, UK: Psychology Press. Retrieved from http://repository.cmu.edu/psychology/1197

Ramirez, R., Palencia-Lefler, M., Giraldo, S., & Vamvakousis, Z. (2015). Musical neurofeedback for treating depression in elderly people. Auditory Cognitive Neuroscience, 9, 354.

Rayner, K. (1975). Parafoveal identification during a fixation in reading. Acta Psychologica, 39, 271–281.

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

Rodden, K., Fu, X., Aula, A., & Spiro, I. (2008). Eye-mouse coordination patterns on web search results pages. In Proceedings of CHI'2008 Extended Abstracts (pp. 2997–3002). New York: ACM.

Ruthven, I. (2014). Relevance behaviour in TREC. Journal of Documentation, 70, 1098–1117.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23, 2507–2517.

Saracevic, T. (1975). RELEVANCE: A review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science and Technology, 26, 321–343.

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. Journal of the American Society for Information Science and Technology, 58, 2126–2144.

Schotter, E.R., Angele, B., & Rayner, K. (2011). Parafoveal processing in reading. Attention, Perception, & Psychophysics, 74, 5–35.

Simola, J., Salojärvi, J., & Kojo, I. (2008). Using hidden Markov model to uncover processing states from eye movements in information search tasks. Cognitive Systems Research, 9, 237–251.

Smucker, M.D., Guo, X.S., & Toulis, A. (2014). Mouse movement during relevance judging: implications for determining user attention. In Proceedings of SIGIR'2014 (pp. 979–982). New York: ACM.

Soleymani, M., Kaltwang, S., & Pantic, M. (2013). Human behavior sensing for tag relevance assessment. In Proceedings of International Conference on Multimedia (pp. 657–660). New York: ACM.

Taylor, A. (2012). User relevance criteria choices and the information search process. Information Processing and Management, 48, 136–153.

Villa, R., & Halvey, M. (2013). Is relevance hard work? Evaluating the effort of making relevant assessments. In Proceedings of SIGIR'2013 (pp. 765–768). New York: ACM.

Voorhees, E.M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In Proceedings of SIGIR'1998 (pp. 315–323). New York: ACM.

Wang, S., Gwizdka, J., & Chaovalitwongse, W.A. (2015). Using wireless EEG signals to assess memory workload in the n-back task. IEEE Transactions on Human–Machine Systems, 46, 424–435.

White, R.W., & Buscher, G. (2012). Text selections as implicit relevance feedback. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1151–1152). New York: ACM.

White, R.W., & Kelly, D. (2006). A study on the effects of personalization and task information on implicit feedback performance. In Proceedings of CIKM'206 (pp. 297–306). New York: ACM.

White, R.W., Ruthven, I., & Jose, J.M. (2002). The use of implicit evidence for relevance feedback in web retrieval. In F. Crestani, M. Girolami, & C. J. van Rijsbergen (Eds.), Advances in information retrieval (pp. 93–109). Berlin: Springer.

Wierda, S.M., Rijn, H.V., Taatgen, N.A., & Martens, S. (2012). Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. PNAS, 109, 8456–8460.

Wilson, D., & Sperber, D. (2002). Relevance theory. In G. Ward & L. Horn (Eds.), Handbook of pragmatics. Oxford, UK: Blackwell.