Exploring Peripheral Physiology as a Predictor of Perceived Relevance in Information Retrieval

Oswald Barral[1], Manuel J. A. Eugster[2], Tuukka Ruotsalo[3], Michiel M. Spapé[3], Ilkka Kosunen[1], Niklas Ravaja[3,4], Samuel Kaski[1,2], Giulio Jacucci[1,3]
Helsinki Institute for Information Technology HIIT
[1]University of Helsinki, Department of Computer Science, PO Box 68, 00014, Finland
[2]Aalto University, Department of Computer Science, PO Box 15400, 00076, Finland
[3]Aalto University, PO Box 15600, 00076, Finland
[4]University of Helsinki, Department of Social Research, PO Box 54, 00014, Finland
first.last@hiit.fi

ABSTRACT
Peripheral physiological signals, as obtained using electrodermal activity and facial electromyography over the corrugator supercilii muscle, are explored as indicators of perceived relevance in information retrieval tasks. An experiment with 40 participants is reported, in which these physiological signals are recorded while participants perform information retrieval tasks. Appropriate feature engineering is defined, and the feature space is explored. The results indicate that features in the window of 4 to 6 seconds after the relevance judgment for electrodermal activity, and from 1 second before to 2 seconds after the relevance judgment for corrugator supercilii activity, are associated with the users' perceived relevance of information items. A classifier verified the predictive power of the features and showed up to 14% improvement predicting relevance. Our research can help the design of intelligent user interfaces for information retrieval that can detect the user's perceived relevance from physiological signals and complement or replace conventional relevance feedback.

Author Keywords
Electrodermal Activity; Corrugator Supercilii; Information Retrieval; Peripheral Physiology; Relevance Prediction; Implicit Relevance Feedback

ACM Classification Keywords
H.3.3. Information Search and Retrieval: Relevance feedback; H.1.2. User/Machine Systems: Human factors; H.5.2. User Interfaces: Evaluation/Methodology

INTRODUCTION
Information retrieval research relies on methods that are able to distinguish relevant from irrelevant information. These methods are based on obtaining relevance assessments from users when they are examining specific information items. The relevance assessments can then be utilized in feedback loops to specify the information need in subsequent iterations [13], direct a search using visual interfaces [24, 29], or simply gather relevance assessments from users for evaluation purposes.

One way to obtain relevance assessments is implicit feedback by monitoring the user, that is, gathering user data in an unobtrusive way while users are engaged with an information retrieval system. Implicit monitoring has been found to be one of the most useful sources for acquiring relevance assessments from the user as it is does not require users to explicitly provide relevance judgments [18]. Previous research has found evidence for implicit behavioral measures, such as dwell-time (the time the user spends to examine an information item) and click-through activity, being useful predictors of perceived relevance [1, 14].

There is rising interest in implicit signals that could provide information about the user's perceived relevance without the requirement to rely on behavioral measures. Physiological signals that can capture users' emotions, attention, and focus could help information retrieval systems determine the relevance of the content for the user without the need to rely on behavioral measures that are dependent on the user interface and interaction design.

This paper reports on an experiment with 40 users, in which the use of two of the most applicable physiological signals, electrodermal activity (skin conductance) and corrugator supercilii activity (brow muscle) were studied as a potential source of implicit signals associated with perceived relevance. These peripheral physiological signals were chosen for this study because they are low cost, unobtrusive, and have been previously associated with psychophysiological functions that could be associated with perceived relevance [23, 30, 31, 33]. The signals were recorded in response to textual content shown during an information retrieval task, their association with perceived relevance was explored, and a classification experiment was conducted to predict relevance for unseen users and information items.

We analyzed the raw signals to study whether the physiological signals could be associated with perceived relevance. Then a set of features was engineered, and a set of important features was selected by exploratory data analysis. Finally, a classifier trained with the chosen features verified the power of the physiological signals for predicting relevance for un-

IUI 2015, March 29–April 1, 2015, Atlanta, GA, USA.

http://dx.doi.org/10.1145/2678025.2701389

seen participants. The main findings of these experiments are the following:

1. The best-suited time windows for relevance prediction from electrodermal activity were found to be from 4 seconds to 6 seconds after the relevance judgment.
2. The best-suited time windows for relevance prediction from corrugator supercilii activity were found to be from 1 second before to 2 seconds after the relevance judgment.
3. A classifier verified the predictive power of the physiological signals with an improvement of 14% against a random baseline.

The rest of the paper is organized as follows. First, we review the background related to relevance feedback and to the use of peripheral physiology in information retrieval settings. Then we describe the experimental study, basic signal analysis, feature engineering, and the results of the exploratory feature analysis. Then we explain the classification setup and characterize the power of the signals for relevance predictions. Finally, we conclude with a discussion and future work. Figure 1 illustrates the analysis workflow, from the data collection to the predictive models.

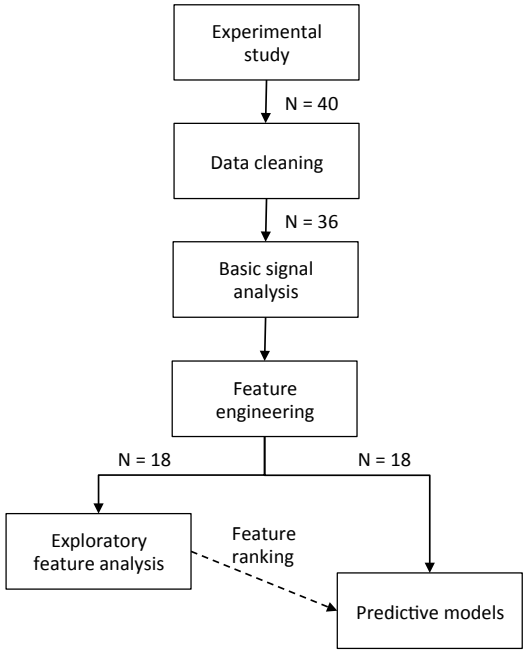

Figure 1. Workflow of the analysis. Physiological data of 40 participants was recorded during the experimental study. Data of four of the participants were rejected due to misplaced or loose sensors. Then, basic signal analysis was carried out in order to examine the association between the physiological signals and the perceived relevance of information items. Then, a set of 25 features was extracted from each of the signals, and the participants were split into two groups of the same size. The first group was used to explore the feature space and generate a feature ranking, which was used to train three predictive models. The models were then tested on the group of unseen participants.

BACKGROUND

Detecting perceived relevance of information presented for users is a central task of interactive information retrieval systems [22]. The perceived relevance can then be used to gather relevance assessments for evaluation purposes [16], direct a search in an interactive loop with the system [25], disambiguate the user's information need [27], or even to measure the user's satisfaction with the information retrieval system [9]. Relevance detection can be based on either explicit or implicit signals and is called explicit feedback and implicit feedback, respectively.

Explicit feedback is a robust method, as it selects relevant and irrelevant information based on direct user interaction [20]. Unfortunately, explicit feedback is operationalized at the expense of users' cognitive resources [16], as the user has to explicitly give commands for the information retrieval system to indicate which information is relevant or irrelevant. Explicit feedback techniques also suffer from a trade-off regarding the user's willingness to invest time to examine the returned information because the relevance feedback can only be targeted to contents that are explicitly judged by users. Eventually, as the task complexity increases the cognitive resources required from the users, the process of relevance assessment may turn into a non-trivial task [20].

Despite its robustness, explicit feedback is therefore often practically insufficient due to the cognitive burden that it causes for the user [17]. Implicit relevance feedback has been proposed to overcome this cognitive burden. The idea of implicit relevance feedback is that relevance of an information item is inferred from interactional data during the user's natural interaction with the search user interface [15].

Implicit feedback [18] has been proposed to obtain relevance assessments by passively observing searchers as they interact with the system. Implicit feedback has been implemented either through the use of surrogate measures based on interaction with documents (such as reading time, scrolling, or document retention) or using other interaction data. Implicit feedback has been shown to have mixed effectiveness because the measures that are good indicators of user interest are often affected by several factors, making the inferences drawn from user interaction not always valid [34].

Affective feedback is a specific type of implicit feedback, and has recently been under active research with a focus on using affective signals to detect the relevance of information [3]. Affective feedback is based on the idea that physiological signals are often associated with cognitive functions relevant to perceiving relevance [4]. Recently, preliminary evidence supporting a combination of basic implicit signals with physiological signals has been reported [21]. One of the main advantages of using physiological signals as implicit input, compared with traditional implicit monitoring, is that these signals can be detected within seconds or even milliseconds after the information items are shown to the user. Physiological signals have been shown to correlate with attention, focus, and semantic memory performance [6, 19], becoming a promising source for acquiring implicit user feedback.

Electrodermal activity (EDA), also known as galvanic skin response (GSR) or skin conductance response (SCR), among others, is a physiological signal that measures the changes in

the electrical properties of the skin, due to the varying level of sweat-induced moisture. EDA has commonly been used to measure the activation of the sympathetic nervous system; therefore, it has proved to be a good indicator of the level of psychological, physiological and emotional arousal [2]. More recently, the short-term (phasic) EDA response has proved to be a useful indicator of stimulus novelty, intensity, emotional content, and significance [23]. Therefore, electrodermal activity has potential to be associated with relevance judgments while it is possible to be captured with very low cost and nonintrusive setup. For example, EDA sensors can be mounted in a computer mouse.

Facial Electromyography (fEMG) is the technique of measuring electrical activity associated with contractions of the facial muscle fibers. One of the muscle groups measured through facial electromyography is the corrugator supercilii muscle (brow muscle). The corrugator supercilii activity (CSA) is a physiological signal particularly promising for determining the relevance of information because it can be used to index negative valence, or frowning, mental workload, fatigue, and compensatory mental effort [30, 31, 33]. Corrugator supercilii activity also increases during tasks requiring heightened effortful attention [6]. As with EDA, it is possible to monitor CSA in a very low cost and nonintrusive setup, as it can be measured even via computer vision.

While some physiological signals have been found to be associated with cognitive functions related to perceived relevance, complete information retrieval systems that make use of physiological computing are still at an early stage. Recent work has addressed implicit inference of relevance using peripheral physiology [5], physiological and affective measures [21], or brain signals [8]. Nevertheless, it is still unclear which signals are the most useful and how they should be used (e.g., what are the best time windows for physiology-based relevance prediction). Our study seeks to address the abovementioned issues by focusing on two signals from the peripheral physiology and exploring their association with perceived relevance. Moreover, we study the predictive power of these signals for automatically detecting relevance.

EXPERIMENTAL STUDY

The present study was designed to 1) investigate how the perceived relevance of users of an information retrieval system is associated with peripheral physiology, 2) study whether perceived relevance can be predicted from features extracted from physiological signals, and 3) determine at what point in time the signals are best suited to indicate the perception of relevance. We recorded electrodermal activity (EDA) and corrugator supercilii activity (CSA) while participants were examining information items returned by a real information retrieval system.

Participants

Forty participants, 34 males and 6 females, participated in the study. We ensured that participants had previous experience in browsing scientific databases and that they were not under psychopharmacological medication. The age of the participants ranged from 21 years old to 47 years old (Mean = 28.17, Median = 26.5). Most of them were post-graduate (37) and the rest were undergraduate students. Only one of the participants reported being a native English speaker, and 17 different mother tongues were reported. Nevertheless, the overall English reading skills were self-reported as advanced. Participants reported themselves to be physically and mentally healthy.

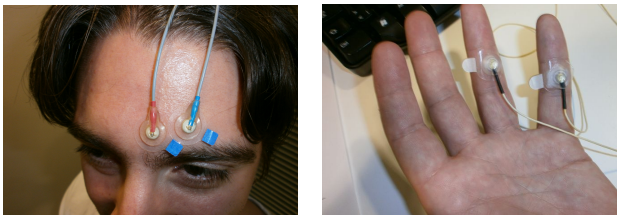

Figure 2. Physiological sensor setup. Sensors were placed in sites overlying the left corrugator supercilii muscle (left) and in the medial phalanges of the participant's left ring and little fingers to measure electrodermal activity (right).

Materials

The setup of the physiological sensors is illustrated in Figure 2. A QuickAmp (BrainProducts GmbH., Germany) amplifier recorded electrodermal activity (EDA) and corrugator supercilii activity (CSA) at a sample rate of 1000 Hz.

CSA sensors were filled with SYNAPSE conductive electrode cream (Kustomer Kinetics Inc., USA) and placed on sites overlying the left corrugator supercilii muscle regions as recommended by Fridlund and Cicoppo [10]. EDA electrodes were filled with TD-246 skin conductance electrode paste (Med Associates Inc., USA) and attached to the middle phalanges of the ring and little fingers of each participant's left hand after her hands were washed with soap and water [7].

The stimulus was presented in the Google Chrome browser. A custom JavaScript code was injected to record the exact time in milliseconds, in relation to the PC clock, when each event took place. To ascertain the synchrony between the browser timestamps and the physiological data, every second the experiment PC sent a synchronization pulse through the parallel port to the QuickAmp amplifier, and each of the browser events was synchronized to the closest pulse.

Task

The task was designed so that the participants could perform an actual search on a real topic of their interest, while still controlling for as many confounding variables as possible, by presenting only one search result at a time in the middle of the screen. The participants were presented with a search box and instructed to perform a query on a topic they were familiar with. They were furthermore informed that they were browsing a scientific database and consequently encouraged to select topics accordingly.

The system presented, in randomized order, six abstract snippets. Of these six snippets, three were always actual search results (relevant to the participant's query) and the other three were randomly generated (irrelevant to the participant's

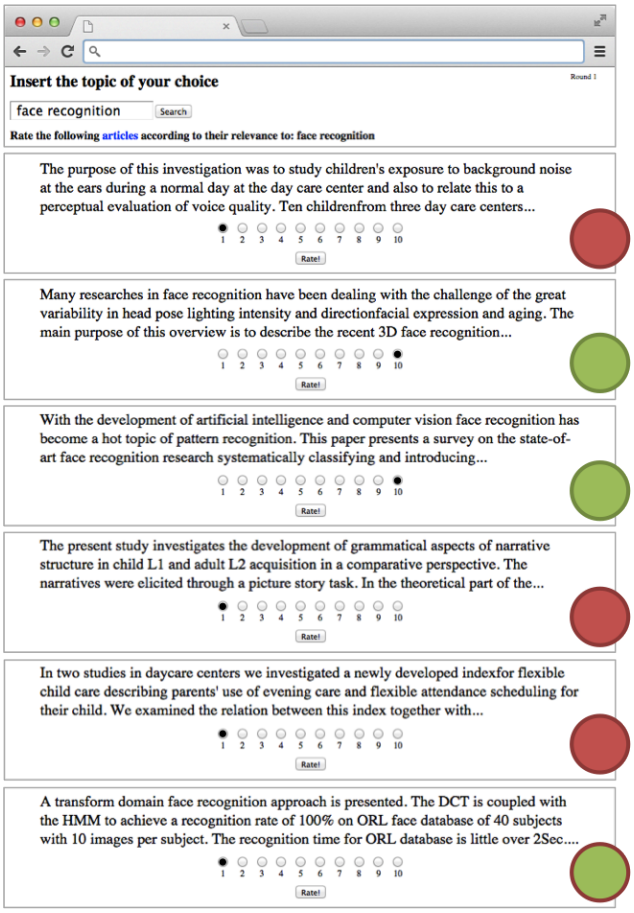

Figure 3. Experimental task and user interface. The participant submits a query, and the system retrieves six abstract snippets of which three are relevant and three are irrelevant. These are presented for the participant one at a time in a randomized order. The participant rates each result using a 1–10 scale. This is repeated for a total of six queries. The figure shows one specific query within a session where the participant searched for "face recognition." The rating scales show the participant judgments. The inner colors of the discs show the ground truth of the abstracts as returned by the search engine (green denotes *relevant*, red denotes *irrelevant*). The outer rings of the discs show the binarized participant judgments. For the first five abstracts, the participant rated according to the ground truth. The last relevant abstract was, according to the participant's judgment, irrelevant.

query). Each abstract was shown until the participant responded by rating the relevance on a scale from 1 to 10, which took ca. 8 seconds on average. Then, the snippet was replaced with the next snippet until all 6 were rated and the participant was asked to perform a search on the next topic. The experiment was completed after the participant rated 36 abstracts (6 topics x 6 abstract snippets).

Search Engine and Content Database

We built a custom search engine and user interface to have full control of the retrieval process, the presentation of results, and the content indexed by the retrieval system. We used a state-of-the-art unigram language modeling approach with Bayesian Dirichlet smoothing to rank the results [36]. The relevant results were retrieved based on the ranking provided by the ranking model directly. The irrelevant results were selected randomly with an additional boolean constraint to exclude results that contained words from the participant's query.

The content items were from a scientific article database consisting of over 50,000,000 articles from the Web of Science prepared by Thomson Reuters, Inc. and from the Digital Libraries of the ACM, the IEEE, and Springer. The first 40 words of the abstracts of the articles were used as result snippets, which we had found in pilot studies to be sufficient for the participants to decide whether the article is relevant.

Procedure

At the beginning of the session, the participants were briefed as to the procedure and purpose of the experiment, before signing informed consent. They were furthermore informed of their right to withdraw from the experiment at any time without any negative consequence. No training session was provided prior to the task, as the interaction with the system was particularly intuitive. Participants could first type a query, such as "face recognition" as shown in Figure 3. Then, the search engine returned six articles, three relevant and three irrelevant. The abstract snippets of the article were then shown for the participant one at the time in a randomized order. The participants then read the abstract snippet and was asked to rate the relevance of the article as soon as they made a decision on the relevance, without the need of reading the text until the end. After rating, the next article was shown. The procedure was repeated for a total of six user-selected topics. After the experiment, the participants were asked to fill in an online survey regarding their background information, and their participation was compensated with two movie tickets.

Data Cleaning

After visual inspection, data from four out of the 40 participants had to be rejected due to loosened or misplaced sensors. The data were binarized to irrelevant and relevant categories (illustrated in Figure 3 by the clicked radio buttons and colored outer rings of the discs). When the relevance judgment explicitly acquired from the participant was less than 4, it was categorized as an irrelevant judgment, and when the relevance judgment was higher than 7, it was categorized as a relevant judgment. Other trials were not categorized, as the judgments were interpreted as ambiguous and the psychophysiological responses associated to them were likely to be misleading.

The physiological signals were first filtered to reduce noise and artifacts. For electrodermal activity we used a low-pass filter with the cut-off at 5 Hz, and for corrugator supercilii activity, a high-pass filter with the cut-off at 10 Hz. To normalize the data, the signals for each participant were then divided by the standard deviation of the signal for that participant.

BASIC SIGNAL ANALYSIS

In order to get an idea on whether there was information about perceived relevance in the physiological signals, we started by analyzing the raw signals. We looked at an 8-second window time-locked to the moment when the participant gave the

explicit judgment. The window spanned from 2 seconds before to 6 seconds after the explicit rating. For every second within this window, we computed the average signal value (i.e., average downsampling), resulting in eight values for each trial.

For both EDA and CSA signals, we executed the following analysis. Within each participant, we aggregated relevant and irrelevant trials with both the arithmetic mean and the more robust median. For a basic overview, we then computed the grand average, i.e., the mean or median of the participant-specific mean or median values for each second of the time window. For a more in-depth analysis, we computed repeated-measures analysis of variance (ANOVA) based on the mean or median values, with two-level factor "relevance" (relevant vs. irrelevant) and eight-level factor "time" (eight seconds of the window).

The results of the computed ANOVAs were corrected using the Greenhouse-Geisser correction on the degrees of freedom, as Mauchly's Test indicated that the assumption of sphericity had been violated.

Results

Figure 4 shows the grand average based on the mean values with 95% confidence intervals for both signals. In case of EDA, a difference was visible between relevant and irrelevant trials around 4 to 6 seconds after the explicit relevance judgment. In the case of CSA, a difference was visible around 1 second after the explicit relevance judgment. The confidence intervals mostly overlap, which is a first indication of the hardness of this prediction problem. The grand average based on the median values showed similar structure.

The ANOVA based on the mean EDA values showed a significant main effect of time, $F(1.46, 51.25) = 20.56, p < 0.0001$. The ANOVA based on median EDA values showed a significant main effect of relevance $F(1.00, 35.00) = 6.19, p < 0.02$ as well as time $F(1.13, 39.40) = 56.44, p < 0.0001$. The ANOVA based on the mean CSA values showed a significant main effect of time $F(4.26, 148.99) = 5.97, p < 0.0001$. In case of the median-based ANOVA, main effect of time on CSA activity was found as well, $F(3.83, 134.04) = 8.99, p < 0.0001$.

For both signals EDA and CSA, main effect of time was found, which indicates that the physiological signal changes reliably due to the relevance judgment. However, the direction of the judgment was only significant for electrodermal activity, as indicated by the significant main effect of the relevance. This means that decision-related physiological changes in corrugator supercilii activity are either not related to perceived relevance, too weakly related to become visible, or not stable enough across time as to cause an interactive effect.

FEATURE ENGINEERING

The results found in the previous section indicate the presence of information on perceived relevance in the psychophysiological responses. However, in order to capture this information, more sophisticated representations of the EDA and CSA signals are needed. The skin conductance response (SCR) elicited by a stimulus can take 1 to 3 or 4 seconds to manifest. After that, electrodermal SCR takes one to three seconds to reach its peak [7]. In contrast to electrodermal activity, electromyography response is fast, as corrugator supercilii activity is found to be elicited at most two seconds after processing a stimulus [35].

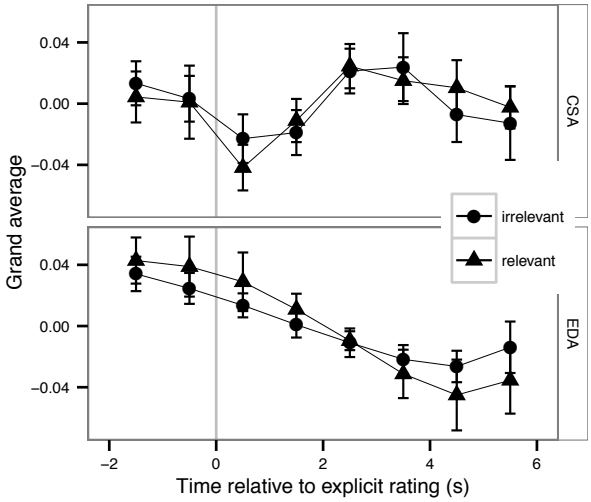

Figure 4. Grand average with 95% confidence interval within the 8-second window of the electrodermal activity (EDA, bottom) and corrugator supercilii activity (CSA, top) signals averaged over participants and trials. The vertical gray line at "0" indicates the explicit rating event. Differences in the signal can be observed around 1 second after the rating for corrugator supercilii activity and around 4 to 6 seconds after the rating for electrodermal activity.

Feature window

Though it is very hard to know exactly when the implicit decision happened, there are two constraints that define the decision moment. First, it is clear that a person needs to read at least a few words to know what the text is about. The reading time required to assess the relevance of a text in our experiment is highly variable (Mean = 8.3s, SD = 4.5s), which is a large obstacle for stimulus-locked analyses (i.e. the window is locked to the moment when the text item appears on the screen). Second, as we instructed the participants to assess the relevance as soon as they had made any decision on it, even though relevance-related processing may start earlier, the final implicit decision on the relevance is expected to occur only shortly before the explicit rating (solid rectangle in Figure 5), which is best captured using response-locked analysis (i.e. the window is locked to the moment when the participant rates the information item). We therefore defined a time window that included a short time before the explicit decision (2s), which was considered to be sufficient to cover the 100s of milliseconds elapsed between the implicit decision and the explicit rating in relation to awareness, preparation of response and response execution (Sternberg stimulus-response model [28]); and a relatively long time following it (6s), as especially EDA changes very slowly (SCR can take up to 6 or 7 seconds to reach its peak [7]).

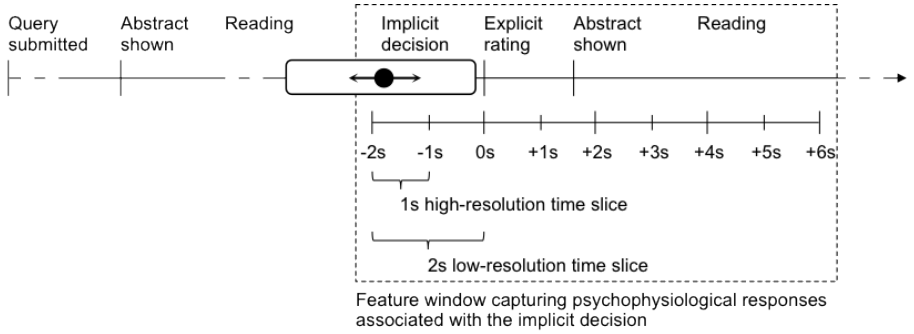

Figure 5. Feature window. The participant first submits the query; then a sequence of abstract snippets is presented, one snippet at a time. For each of those, the participant reads the text for an unfixed amount of time ($M = 8.3, SD = 4.5$, in seconds) and, after making a decision regarding its relevance to the query, rates accordingly. The time window selected to generate the features is outlined in the figure; it goes from 2 seconds before to 6 seconds after the rating. The decision on the relevance was assumed to happen at most 2 seconds before the rating, and the feature window was selected to capture the psychophysiological responses associated with it, which for corrugator supercilii activity takes around one to two seconds and in the case of electrodermal activity can take up to 6 or 7 seconds. Features were generated from eight 1-second (high-resolution) and four 2-second (low-resolution) time slices of the feature window.

Even though the selected time window overlaps with the next item presentation, the randomization of the presentation order of relevant and irrelevant text items together with the large number of participants should redress confounding noise and artifacts introduced by the next presented information item.

Feature generation

We generated features within this 8-second time window (dashed rectangle in Figure 5), time-locked to the moment when the participant gave the explicit rating ("0s" in Figure 5). Given the psychophysiolgical literature and the experimental procedure, it is now likely that this window contains the psychophysiological response associated with the point in time when the participant processed the stimuli and made a decision whether the information is relevant or irrelevant (i.e., implicit decision, black dot in Figure 5). Within the window, we generated high-resolution and low-resolution features; the first ones were based on nonoverlapping 1-second slices, the latter based on nonoverlapping 2-second slices.

Using this basic structure, we generated features describing four different characterizations of the signals within the time window of a trial. Let h indicate high-resolution features and l indicate low-resolution features. The set $I_h = \{-2, -1, 0, 1, 2, 3, 4, 5\}$ describes the start seconds of the 1-second slices, the set $I_l = \{-2, 0, 2, 4\}$ describes the start seconds of the 2-second slices, and the set $I_d = \{-1, 0, 1, 2, 3, 4, 5\}$ describes the middle second of adjacent 1-second slices. Furthermore, $s^{i,j}$ describes the values of the physiological signal $s \in \{\text{EDA}, \text{CSA}\}$ from i seconds to j seconds.

The *average signal features* quantify the mean signal per time slice. They are defined as follows:

$$h^i = \text{mean}(s^{i,i+1}) - \text{bl}_h, i \in I_h$$
$$l^j = \text{mean}(s^{j,j+1}) - \text{bl}_l, j \in I_l$$

These features are baseline-corrected, that is, centered by subtracting a trial-based baseline bl_h and bl_l.

The *difference features* quantify the amount of change of the signals between adjacent time slices. They are defined as followed:

$$d^k = (\text{mean}(s^{k,k+1}) - \text{mean}(s^{k-1,k})) - \text{bl}_d, k \in I_d$$

These are baseline-corrected, as they are centered by a trial-specific baseline bl_d describing the mean differences.

The *maximum signal features* quantify the maximum signal per window. They are defined as follows:

$$\text{max}_h = \max(h^i, i \in I_h)$$
$$\text{max}_l = \max(l^j, j \in I_l)$$
$$\text{max}_d = \max(d^k, k \in I_d)$$

The *latency features* quantify the latency to the maximum value. They are defined as follows:

$$\text{lat}_h = \text{argmax}_i(h^i, i \in I_h)$$
$$\text{lat}_l = \text{argmax}_j(l^j, j \in I_l)$$
$$\text{lat}_d = \text{argmax}_k(d^k, k \in I_d)$$

The defined characterizations describe a general feature engineering framework for physiological signals, where each signal is described by 25 features. Table 1 shows the final set of features generated for EDA and CSA using this procedure.

EXPLORATORY FEATURE ANALYSIS

In order to analyze the quality and importance of the individual features, we conducted an exploratory feature analysis. We wanted to identify specific features that are associated with perceived relevance using data of half of the valid participants ($N = 18$). We ranked the features according to their generalizability across participants utilizing a widely used feature-selection technique based on the filter principle [26]. Electrodermal activity and corrugator supercilii activity features were ranked separately using identical procedures.

	Electrodermal activity (EDA)					Corrugator supercilii activity (CSA)				
	Features	#Part.	$\bar{W}$	SD(W)	Rank	Features	#Part.	$\bar{W}$	SD(W)	Rank
Model M_1	lat_d	9	158	38	1	d^0	7	163	42	1
	d^5	7	170	50	2	h^0	7	159	57	
	d^3	7	151	38		max_h	7	141	36	
	d^2	7	149	33		l^0	6	156	43	2
	h^5	7	146	36		h^{-2}	5	153	29	3
	l^4	7	146	24		max_d	5	137	48	
	d^4	7	140	47		d^2	4	153	23	4
Model M_2	h^4	6	146	25	3	h^{-1}	4	136	16	
	l^2	4	133	44	4	h^4	4	130	39	
	max_d	4	122	29		d^{-1}	4	130	40	
Model M_3	h^3	3	156	2	5	lat_d	4	123	21	
	h^{-2}	3	148	39		h^3	4	117	43	
	max_l	3	140	33		l^{-2}	3	160	41	5
	h^2	3	125	49		d^5	3	153	14	
	l^{-2}	2	168	36	6	lat_l	3	150	20	
	max_h	2	144	36		d^4	3	140	6	
	d^1	2	123	44		l^2	3	133	38	
	d^{-1}	2	110	23		max_l	3	123	53	
	l^0	2	81	4		l^4	2	145	21	6
	h^0	2	75	11		d^1	2	135	25	
	lat_l	1	156	–	7	h^1	2	130	34	
	lat_h	1	119	–		lat_h	2	126	30	
	h^{-1}	1	111	–		h^2	2	122	52	
	h^1	1	78	–		d^3	1	136	–	7
	d^0	1	77	–		h^5	0	–	–	8

Table 1. The features were named as follows: h^i are the features generated in the high-resolution time window, from second i to second $i+1$, relative to the moment the participant gives explicit feedback. The l^i are the features generated in the low-resolution time window from second i to second $i+2$, relative to the moment the participant gives explicit feedback. The d^i are the features generated from the difference between i to $i+1$ and $i-1$ to i seconds relative to the moment the participant gives explicit feedback. $max_{h,l,d}$ and $lat_{h,l,d}$ are generated from the maximum value and latency to the maximum value in the high-resolution (h), low-resolution (l), and difference (d) features. Electrodermal activity-derived features are shown in the left part of the table and corrugator supercilii activity-derived features in the right. *#Part.* indicates the number of participants for whom the feature appears in the *top features*. The mean of the W statistic and the standard deviation for these participants are indicated in columns $\bar{W}$ and *SD* respectively. The features are ranked according to the number of times they are included in a participant's *top features*, which is indicated by *Rank*. Additionally, the set of features used for each of the three predictive models M_1, M_2, and M_3 is indicated as well.

For each participant, the Wilcoxon rank-sum statistic (W) was computed for every feature between relevant and irrelevant trials. The five features with highest W including draws were then selected for each participant (i.e., *top features*). Following, in order to analyze the features over participants, we ranked the features according to the number of times the features were included in a participant's *top features*.

Results

Table 1 contains the ranked features for electrodermal activity and corrugator supercilii activity, according to the number of participants' *top features* each feature belongs to. The mean W and standard deviation for these participants are included as well in the table.

Electrodermal activity was found to have the highest association with relevance via the lat_d feature, as half of the participants had this feature in their *top features*. This feature represents the latency to the maximal difference between two adjacent slices. Therefore, the point in time in which the maximum increase of EDA occurs appears to be a good indicator of the content of the relevance judgments. Interestingly, h^5, l^4, d^5, and d^4 are some of the top-ranked features and are all generated from the time window from 4 to 6 seconds after the relevance judgment. The h^5 and l^4 features refer to the amount of electrodermal activity between 5 to 6, and 4 to 6 seconds after the relevance judgment is made, respectively. The d^5 feature is a measure of how much the electrodermal activity changes five seconds after the relevance judgment, and d^4 measures the change of EDA around 4 seconds after the explicit judgment. As previously pointed out, the skin conductance response (SCR) elicited by a stimulus can take up to 6 or 7 seconds to reach its peak [7]. Thus, considering the implicit decision moment as the trigger for SCR, these features could well be related to the SCR peak, which in turn would be associated with relevance judgments. Moreover, the mean W for d^5 is very high compared with the other features. This is an indicator that for the participants for whom this feature is in the *top features*, there is a strong association between the increase of EDA around 5 seconds after the judgment, and the content of the judgment. Finally, it is worthy to recall the fact that many *difference features* appear in the top of the table for electrodermal activity, as five out of the seven first- and second-ranked features are *difference features*. This fact points in the direction that the association of EDA signal with perceived relevance relies on how the signal changes across time, more than in absolute values.

Corrugator supercilii activity was associated with relevance via the features generated from the signal around the moment when the participant made explicit the relevance judgment. The d^0 feature refers to the change in CSA between the slice before and the slice after the explicit relevance judgment is made. The h^0 and l^0 features refer to the first 1-second and 2-second slice after the relevance judgment, respectively. Based

on the results of the ranking, it is likely that at least these three features, and therefore the time window from 1 second before to 2 seconds after the relevance judgment, are associated with users' perception of relevance of information items. As previously pointed out, corrugator supercilii activity is found to be elicited at most 2 seconds after processing a stimulus. Therefore, the finding that the highest ranked features are in this time window is coherent with the psychophysiological literature. Hence we can deduce that in the instants immediately before and after the relevance judgment is made, corrugator supercilii activity is potentially associated with perceived relevance. It is also interesting to point out that the maximum 1-second slice in the signal (i.e., max_h) belongs to the group of first-ranked features. Therefore, it appears that, contrarily to electrodermal activity where the association with perceived relevance seems to come from the changes in the signal (*difference features*), for corrugator supercilii activity, the peak value in the signal seems to be one of the properties more strongly associated with perceived relevance, together with the signal values around the moment of the explicit rating.

PREDICTIVE POWER

In order to verify that our generated features have predictive power, we built classifiers that predict relevance judgments for the 18 participants of our study not used in the exploratory feature analysis. This allowed us to test the generalizability of the features. We used a multiview learning method, a leave-one-participant-out strategy, and assumed the relevance judgments to be balanced [8, used the same scheme]. The latter assumption reassembles our original experimental design and allowed us to focus on the predictive power of the features rather then the problem of potentially imbalanced relevance judgments.

Prediction model

In detail, we used a multiple kernel learning (MKL) method to learn classification models of the following form:

$$y = f(\mathbf{v}_1, \ldots, \mathbf{v}_K) = \sum_{k=1}^{K} \beta_k \langle \mathbf{w}_k, \Phi_k(\mathbf{v}_k) \rangle + b.$$

Here, y denotes the binary relevance judgment, and $\mathbf{v}_k$ are the features generated from the different signals (called a view) that is, $\mathbf{v}_1$ is the view with the features for EDA and $\mathbf{v}_2$ is the view with the features for CSA. $\langle \cdot, \cdot \rangle$ denotes the scalar product, $\mathbf{w}_k$ the weight vector of the observations, $\Phi_k(\mathbf{v}_k)$ the feature map of the view $\mathbf{v}_k$, β_k the kernel weights, and b the bias. Given the selected features for each view $\mathbf{v}_k$, we normalized the data and computed a Gaussian kernel with the kernel width defined as the median distance between the observations [12]. For the concrete estimation of the classification models, we use a Bayesian MKL algorithm with an efficient inference based on variational approximation [11].

Prediction setup

We applied a leave-one-participant-out learning strategy as follows. For each participant we learned a classification model using the other participants' data (i.e., the remaining seventeen participants). The prediction accuracy was then computed on the participant's relevance observations. The number of relevance observations varies slightly for each participant because of the binarization. We established balanced data by randomly drawing the learning set and the test set from the set of relevant and the set of irrelevant observations, each with the number of observations defined by the smaller set. This reassembles our original experimental design and is a simple but well-established strategy to exclude possible problems of the classification method with imbalanced classes. To eliminate a possible observation sampling bias we repeated this procedure five times.

Model	Mean accuracy	SD accuracy	p-value	Mean improvement
M_3	0.5359	0.0814	0.0783	7.19%
M_1	0.5376	0.0787	0.0587	7.51%
M_2	0.5711	0.0997	**0.0076**	14.22%

Table 2. Classification results based on the unseen 18 participants for models based on different feature sets. The table lists the mean classification accuracy, the p-value indicating a significant different mean classification accuracy compared to the random baseline, and the corresponding mean improvement. Because of our experimental design, the random baseline prediction of whether an abstract is relevant or irrelevant is 0.5. Bold entries denote that improvements are statistically significant at a level of $\alpha = 0.05$, p-value $< \alpha$ with correction for multiple testing.

Model definition and feature sets

Based on the exploratory feature analysis, we defined three basic models, M_1, M_2, and M_3 with increasing sizes of feature sets, and computed classification models based on these feature sets. Model M_1 contained the top five features including draws, model M_2 the top ten features including draws, and model M_3 all features from both signals (see Table 1, column #Part.).

Results

Table 2 summarizes the classification accuracies for the models M_1, M_2, and M_3 based on the different feature sets. We report the mean classification accuracy, improvement over the random baseline, and the p-value of a t-test for significance corrected for multiple testing using the Bonferroni correction. The t-test was applicable because the Shapiro-Wilk test did not reject the null hypothesis that the samples come from a normal distribution.

The results verified that there is predictive power in the set of features generated for EDA and CSA and that the features generalize to new users. The M_2 (medium feature set) classification models predicted relevant and irrelevant abstracts for an unseen participant significantly better than the random baseline and achieved a mean improvement of 14%. In the cases of M_1 (smallest feature set) and M_3 (all features) classification models, no significant improvement could be found. Therefore, the M_1 feature set contained insufficient information to discriminate the relevant and irrelevant items. On the other hand, the M_3 feature set contained all possible information, but also introduced a significant amount of noise.

Figure 6 shows the individual classification performances using model M_2 for each of the 18 participants. Points indi-

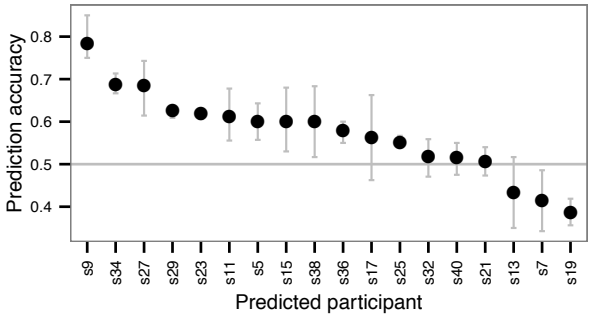

Figure 6. Individual classification accuracy using model M_2 for each of the 18 participants based on training on data of the remaining participants and ordered according to the accuracy. Points indicate the mean accuracy, the error bars show the bootstrap confidence intervals.

cate the mean predicting accuracy, and error bars show bootstrap confidence intervals. The horizontal line at 0.5 marks the random baseline. Fifteen participants are above, and 3 participants are below the random baseline. We also can observe that for some participants, the variation is very small (e.g., participants `s34`, `s29`, and `s23`), whereas for others it is substantial (e.g., participants `s15` and `s38`).

As a final observation, we want to point out participant `s34`, for whom the prediction worked well and with low variation. The exemplary "face recognition" search session in Figure 3 was part of his experiment. The classifier predicted the first five articles according to the participant's explicit ratings. The last article was predicted to be relevant, contrary to the participant's rating. However, the abstract can be seen as relevant. The participant perhaps made a mistake, and we could detect it using physiolgical signals. Even though this is only one specific case, it nicely illustrates the potential of this approach.

DISCUSSION

The results indicate that the EDA and CSA are associated with perceived relevance and the features extracted from the signals transfer to improved accuracy in a classification setup. However, our experimental setup leaves room for further research in three respects.

First, our models in general significantly outperform the random baseline, which indicates that psychophysiology alone can help predict relevance. However, the classification accuracy achieved using only the physiological signals hints that these signals alone are unlikely to serve as relevance predictors. Therefore, psychophysiology is envisioned to be used in combination with other implicit relevance feedback techniques, such as brain signals [8] or facial expressions [5], to strengthen current information retrieval systems.

Second, our experimental design is balanced between relevant and irrelevant abstracts in order to ensure that we measure signals and effects related to relevance judgments. In a real information retrieval setting, however, it is likely that the two classes are imbalanced with the majority of the information items being irrelevant. Experiments with more realistic data and larger amount of observations are needed to show how our results generalize to such scenarios.

Third, the prediction accuracy varies substantially across and within participants. As can be glanced from the left side of Figure 6, our approach achieves good accuracy for about 10 out of 18 unseen participants. Along with the specific observation regarding participant `s34`, this underlines the possibility that there may be substantial differences between users' physiological signals associated with perceived relevance. Other factors need to be taken into account as, for instance, it has been shown that there is a portion of users that are "EDA non-responders" [32]. Consequently, personalized models that are built for each user separately could further improve and stabilize the prediction accuracy.

CONCLUSIONS

Physiological sensors are becoming more ubiquitous and available for every day use. This has raised an interest in study of their usefulness for a wide spectrum of computing applications. We studied peripheral physiological signals for predicting users' perceived relevance when they are engaged in an information retrieval task. In the present work, we concentrated on two of the most promising physiological signals: electrodermal activity (EDA) and corrugator supercilii activity (CSA). Our results suggest that peripheral physiology can be used to predict relevance of information for a user when engaged in an information retrieval task, but that the prediction is sensitive to selecting correct features and time windows. Features in the window of 4 to 6 seconds after the relevance judgment for electrodermal activity (EDA) and from 1 second before to 2 seconds after the relevance judgment for corrugator supercilii activity (CSA) were found to be associated with the perceived relevance. Our findings can help to build systems that can detect relevance from physiology and open a horizon for adaptive intelligent systems that can proactively, with minimum user intervention, react to a user's information needs.

ACKNOWLEDGMENTS

This work has been partly supported by the Academy of Finland (278090, Multivire, 255725; and the Finnish Centre of Excellence in Computational Inference Research COIN, 251170), Re:Know funded by TEKES, and MindSee (FP7 – ICT; Grant Agreement # 611570). Certain data included herein are derived from the Web of Science prepared by THOMSON REUTERS, Inc., Philadelphia, Pennsylvania, USA: Copyright THOMSON REUTERS, 2011. All rights reserved. Data is also included from the Digital Libraries of the ACM, IEEE, and Springer.

REFERENCES

1. Agichtein, E., Brill, E., and Dumais, S. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, ACM (New York, NY, USA, 2006), 19–26.
2. Andreassi, J. L. *Psychophysiology: Human behavior & physiological response*. Psychology Press, 2000.

3. Arapakis, I. Affective feedback: An investigation into the role of emotions in the information seeking process. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, ACM (New York, NY, USA, 2008), 891–891.
4. Arapakis, I., Athanasakos, K., and Jose, J. M. A comparison of general vs personalised affective models for the prediction of topical relevance. In *Proceedings of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, ACM (New York, NY, USA, 2010), 371–378.
5. Arapakis, I., Konstas, I., and Jose, J. M. Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, ACM (New York, NY, USA, 2009), 461–470.
6. Cohen, B. H., Davidson, R. J., Senulis, J. A., Saron, C. D., and Weisman, D. R. Muscle tension patterns during auditory attention. *Biological Psychology 33*, 2-3 (1992), 133 – 156.
7. Dawson, M. E., Schell, A. M., Filion, D. L., and Berntson, G. G. The electrodermal system. In *Handbook of Psychophysiology*, J. T. Cacioppo, L. G. Tassinary, and G. Berntson, Eds., third ed. Cambridge University Press, 2007, 157–181. Cambridge Books Online.
8. Eugster, M. J., Ruotsalo, T., Spapé, M. M., Kosunen, I., Barral, O., Ravaja, N., Jacucci, G., and Kaski, S. Predicting term-relevance from brain signals. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, ACM (New York, NY, USA, 2014), 425–434.
9. Fox, S., Karnawat, K., Mydland, M., Dumais, S., and White, T. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS) 23*, 2 (Apr. 2005), 147–168.
10. Fridlund, A. J., and Cacioppo, J. T. Guidelines for human electromyographic research. *Psychophysiology 23*, 5 (1986), 567–589.
11. Gönen, M. Bayesian efficient multiple kernel learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, J. Langford and J. Pineau, Eds., ACM (New York, NY, USA, 2012), 1–8.
12. Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research 13* (2012), 723–773.
13. Harman, D. Relevance feedback revisited. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '92, ACM (New York, NY, USA, 1992), 1–10.
14. Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, ACM (New York, NY, USA, 2005), 154–161.
15. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., and Gay, G. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *Transactions on Information Systems (TOIS) 25*, 2 (Apr. 2007).
16. Kelly, D. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval 3*, 1-2 (Jan. 2009), 1–224.
17. Kelly, D., and Fu, X. Elicitation of term relevance feedback: An investigation of term source and context. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, ACM (New York, NY, USA, 2006), 453–460.
18. Kelly, D., and Teevan, J. Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum 37*, 2 (Sept. 2003), 18–28.
19. Klimesch, W. Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain research reviews 29*, 2 (1999), 169–195.
20. Koenemann, J., and Belkin, N. J. A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '96, ACM (New York, NY, USA, 1996), 205–212.
21. Moshfeghi, Y., and Jose, J. M. An effective implicit relevance feedback technique using affective, physiological and behavioural features. In *Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, ACM (New York, NY, USA, 2013), 133–142.
22. Moshfeghi, Y., Pinto, L., Pollick, F., and Jose, J. Understanding relevance: An fmri study. In *Advances in Information Retrieval*, P. Serdyukov, P. Braslavski, S. Kuznetsov, J. Kamps, S. Rger, E. Agichtein, I. Segalovich, and E. Yilmaz, Eds., vol. 7814 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013, 14–25.
23. Ravaja, N. Contributions of psychophysiology to media research: Review and recommendations. *Media Psychology 6*, 2 (2004), 193–235.
24. Ruotsalo, T., Jacucci, G., Myllymäki, P., and Kaski, S. Interactive intent modeling: Information discovery beyond search. *Commun. ACM 58*, 1 (Dec. 2014), 86–92.

25. Ruotsalo, T., Peltonen, J., Eugster, M. J., Glowacka, D., Konyushkova, K., Athukorala, K., Kosunen, I., Reijonen, A., Myllymäki, P., Jacucci, G., and Kaski, S. Directing exploratory search with interactive intent modeling. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, CIKM '13, ACM (New York, NY, USA, 2013), 1759–1764.

26. Saeys, Y., Inza, I., and Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics 23*, 19 (2007), 2507–2517.

27. Shen, X., Tan, B., and Zhai, C. Implicit user modeling for personalized search. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, ACM (New York, NY, USA, 2005), 824–831.

28. Sternberg, S. The discovery of processing stages: Extensions of donders' method. *Acta Psychologica 30*, 0 (1969), 276 – 315.

29. Teevan, J., Alvarado, C., Ackerman, M. S., and Karger, D. R. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, ACM (New York, NY, USA, 2004), 415–422.

30. Van Boxtel, A., and Jessurun, M. Amplitude and bilateral coherency of facial and jaw-elevator emg activity as an index of effort during a two-choice serial reaction task. *Psychophysiology 30*, 6 (1993), 589–604.

31. Veldhuizen, I., Gaillard, A., and De Vries, J. The influence of mental fatigue on facial emg activity during a simulated workday. *Biological Psychology 63*, 1 (2003), 59–78.

32. Venables, P., and Mitchell, D. The effects of age, sex and time of testing on skin conductance activity. *Biological Psychology 43*, 2 (1996), 87 – 101.

33. Waterink, W., and Van Boxtel, A. Facial and jaw-elevator emg activity in relation to changes in performance level during a sustained information processing task. *Biological Psychology 37*, 3 (1994), 183–198.

34. White, R. W., Ruthven, I., and Jose, J. M. A study of factors affecting the utility of implicit relevance feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, ACM (New York, NY, USA, 2005), 35–42.

35. Winkielman, P., and Cacioppo, J. T. Mind at ease puts a smile on the face: psychophysiological evidence that processing facilitation elicits positive affect. *Journal of Personality and Social Psychology 81*, 6 (2001), 989.

36. Zhai, C., and Lafferty, J. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS) 22*, 2 (Apr. 2004), 179–214.