Jacucci, Giulio; Barral, Oswald; Daee, Pedram; Wenzel, Markus; Serim, Baris; Ruotsalo, Tuukka; Pluchino, Patrik; Freeman, Jonathan; Gamberini, Luciano; Kaski, Samuel; Blankertz, Benjamin

Integrating neurophysiologic relevance feedback in intent modeling for information retrieval

Published in:
JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY

DOI:
10.1002/asi.24161

Published: 12/03/2019

Document Version
Publisher's PDF, also known as Version of record

Integrating Neurophysiologic Relevance Feedback in Intent Modeling for Information Retrieval

Giulio Jacucci
Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, P.O. Box 68, (Pietari Kalmin katu 5), Helsinki FI-00014, Finland. E-mail: giulio.jacucci@helsinki.fi

Oswald Barral
Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, P.O. Box 68, (Pietari Kalmin katu 5), Helsinki FI-00014, Finland. E-mail: oswald.barral@helsinki.fi

Pedram Daee
Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, P.O.Box 15400, Aalto FI-00076, Finland. E-mail: pedram.daee@aalto.fi

Markus Wenzel
Neurotechnology Group, Technische Universität Berlin, Berlin 10587, Germany. E-mail: markus.wenzel@hhi.fraunhofer.de

Baris Serim
Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, P.O. Box 68, (Pietari Kalmin katu 5), Helsinki FI-00014, Finland. E-mail: baris.serim@helsinki.fi

Tuukka Ruotsalo
Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, P.O. Box 68, (Pietari Kalmin katu 5), Helsinki FI-00014, Finland. E-mail: tuukka.ruotsalo@helsinki.fi

Patrik Pluchino
Human Inspired Technology Research Centre, University of Padova, Via Luzzatti 4, Padova 35121, Italy. E-mail: patrik.pluchino@unipd.it

Jonathan Freeman
Goldsmiths, University of London, New Cross, London SE14 6NW, UK. E-mail: j.freeman@gold.ac.uk

Luciano Gamberini
Human Inspired Technology Research Centre, University of Padova, Via Luzzatti 4, Padova 35121, Italy. E-mail: luciano.gamberini@unipd.it

Samuel Kaski
Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, P.O.Box 15400, Aalto FI-00076, Finland. E-mail: samuel.kaski@aalto.fi

Benjamin Blankertz
Neurotechnology Group, Technische Universität Berlin, Berlin 10587, Germany. E-mail: benjamin.blankertz@tu-berlin.de

Received November 20, 2017; revised April 20, 2018; accepted October 17, 2018

 • Published online Month 00, 2018 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.24161

The use of implicit relevance feedback from neurophysiology could deliver effortless information retrieval. However, both computing neurophysiologic responses and retrieving documents are characterized by uncertainty because of noisy signals and incomplete or inconsistent representations of the data. We present the first-of-its-kind, fully integrated information retrieval system that makes use of online implicit relevance feedback generated from brain activity as measured through electroencephalography (EEG), and eye movements. The findings of the evaluation experiment (N = 16) show that we are able to compute online neurophysiology-based relevance feedback with performance significantly better than chance in complex data domains and realistic search tasks. We contribute by demonstrating how to integrate in interactive intent modeling this inherently noisy implicit relevance feedback combined with scarce explicit feedback. Although experimental measures of task performance did not allow us to demonstrate how the classification outcomes translated into search task performance, the experiment proved that our approach is able to generate relevance feedback from brain signals and eye movements in a realistic scenario, thus providing promising implications for future work in neuroadaptive information retrieval (IR).

Introduction

Information retrieval systems are confronted with a difficult task; deriving a user's information needs from limited explicit user signals and use these to retrieve information matching those needs. Although modeling the data to be retrieved has witnessed dramatic advances during the last decades, understanding users' information needs is still based on rather simple user signals, such as queries, clicks, speech commands, or other explicit interactions. As a result, understanding information needs implicitly without disrupting the user has become a central research challenge in information retrieval (IR). Neurophysiologic measures are promising candidates for implicitly gathering relevance feedback, as they reflect the inner state of the user and can be collected unobtrusively at high throughput (Cowley et al., 2016; Eugster et al., 2016; Jacucci, Fairclough, & Solovey, 2015; Wenzel, Bogojeski, & Blankertz, 2017). Neurophysiologic signals hold a great potential for information retrieval as they provide a novel user signal revealing interests and relevance towards a diverse digital content as they happen when users are consuming digital information. Neurophysiologic signals also carry extraordinary practical promise as numerous types of wearable devices are rapidly becoming integral part of people's everyday life.

However, successful application of neurophysiologic measures in IR encounters a dual uncertainty problem: (a) noisiness and unknown causes of responses in neurophysiologic signals make it difficult to interpret them, a problem exacerbated by the lack of stimulus control in realistic settings, and (b) the IR process involves inherent uncertainty originating from the ambiguity and inconsistency of the representations of data to be retrieved. Unlike explicit relevance feedback that has low uncertainty due a user's overt control, implicit relevance feedback techniques are intrinsically noisy. When observing a user's clickthrough activity or brain responses to infer relevance feedback, the uncertainty of the feedback accuracies becomes higher, and incorporating this feedback within an interactive IR system requires novel computational solutions. The integration of brain signals has been especially challenging; even though they have shown promise, their use beyond laboratory experiments with very controlled stimuli remains largely unexplored. Previous work displays a limited number of unambiguous stimuli on the screen and/or constrains user interaction to decrease the amount of noise (Eugster et al., 2014; Eugster et al., 2016). In contrast, realistic search interfaces are characterized by dense information, potential ambiguity regarding the relevance of search results, and user interaction.

Our work provides the following contributions:

1. We demonstrate an approach able to predict implicit relevance feedback from human-brain measurements in a realistic search scenario.
2. We present a first-of-its-kind interactive IR system that combines brain-based feedback and eye tracking with scarce explicit feedback for improved relevance predictions.

The article is structured as follows. First, a brief discussion on related work on implicit relevance feedback in IR using brain–computer interfaces (BCIs) is presented. The section *An Approach for Single-Trial Relevance Computation in IR* investigates the challenge of decoding single-trial event-related potentials (ERP) that involve semantic interpretation of complex stimuli with large variability. We follow with a detailed proposal of a neurophysiologic approach for relevance computation, providing validation proof for the method, while highlighting potential challenges to be addressed when integrating relevance computation from brain signals in an IR system.

In the subsequent section, *Addressing Uncertainty in an Online Neuroadaptive System through Interactive Intent Modeling* we propose interactive intent modeling as a particular retrieval and ranking approach that facilitates the elicitation of explicit and implicit relevance feedback. Our approach in this respect is characterized by combining modeling of neurophysiologic response with modeling interactively intent in IR. In the section *An Experiment in Neuroadaptive Literature Search* we report the evaluation of our approach through findings from an experiment (N = 16) showing that we can predict neurophysiology-based relevance feedback in complex data domains and realistic search tasks and combine it with explicit relevance feedback in interactive intent modeling.

Related Work

Traditional relevance feedback techniques involve asking a user to provide explicit judgments on the information content. These has proven to be problematic because, in practice, users are reluctant to interrupt their search task to

DOI: 10.1002/asi

provide relevance feedback, even although they are aware that doing so would improve their search performance (Kelly & Fu, 2006). An important bottleneck of information seeking systems is that a considerable amount of user relevance feedback on retrieved items is needed to properly explore the large information space (Daee, Pyykkö, Glowacka, & Kaski, 2016). To overcome this challenge, previous approaches investigated "implicit relevance feedback" as indexed from search behavior from mouse and keyboard interaction data to understand a user's interests and personalize and rank search results (Kelly & Teevan, 2003). Other sources of implicit feedback include eye tracking to infer a user's interest through various metrics such as fixation count, dwell time, pupil size, and scan paths (e.g., Gwizdka, 2014; Oliveira, Aula, & Russell, 2009; Puolamäki, Salojärvi, Savia, Simola, & Kaski, 2005), analysis of user's facial expressions (e.g., Arapakis, Athanasakos, & Jose, 2010), physiologic responses (e.g., Barral et al., 2015, 2016), or a combination of these (e.g., Arapakis, Konstas, & Jose, 2009; Moshfeghi & Jose, 2013). Lately, brain signals have been identified as promising sources for implicit relevance feedback and information personalization (e.g., Eugster et al., 2014; Eugster et al., 2016; Golenia, Wenzel, & Blankertz, 2015).

IR is one of the fields that could profit from this direct access to the mental processes of the brain (Golenia et al., 2015; Gwizdka & Mostafa, 2015, 2017). First of all, mental processes can reveal information about relevance in response to particular information items thereby providing an effective way to elicit implicitly relevance feedback with great efficiency gains in being able to expose more items and collect relevance feedback without disrupting the user's search process. Second, mental processes and psychophysiologic states can be used to automatically annotate information such as news with affective or relevance response for future use and collaborative filtering (Barral et al., 2016; Barral, Kosunen, & Jacucci, 2017), and finally affective states as detectable from the brain can provide important context information for when or how to present information to user considering awareness, cognitive workload and other mental states. Research at the intersection between brain-computer interfaces (BCIs) and IR is still in an early stage, and appropriate neurophysiologic methods have to be matched with the appropriate paradigms for HCI in IR. Kauppi et al. (2015) studied magnetoencephalographic signals alone and in conjunction with gaze signals to provide relevance feedback in an image retrieval task by using a static image database. Similarly, Eugster et al. (2014) decoded the EEG with the objective of providing relevance feedback in a text retrieval task by using a static text data set. Other studies (Golenia et al., 2015; Golenia, Wenzel, Bogojeski, & Blankertz, 2018) demonstrated how the brain response to relevant versus irrelevant information can be harnessed to improve image searches in ambiguous search tasks. Recently research in the neurophysiologic correlates of relevance have been studied by Moshfeghi, Pinto, Pollick, & Jose, (Moshfeghi, Pinto, Pollick, & Jose, 2013) using functional Magnetic Resonance Imaging (fMRI) revealing three brain regions in the frontal, parietal and temporal cortex where brain activity differed between processing relevant and non-relevant documents. Not only where in the brain but also when relevance assessment phenomena happen have been studied for example by Allegretti et al. (2015) using a 64-channel EEG device. They found a significant variation between relevance and nonrelevance for the first 800 milliseconds (ms) of a relevance assessment process from the presentation of the image within the EEG signals. These studies are important as provide important additional evidence on the feasibility to include brain signal based relevance elicitation, however most studies focus on relevance of images and videos and less on text for which ca be more challenging to elicit and detect physiologic responses. Moreover, Eugster et al. (2016) gave relevant feedback on words from the Wikipedia database according to information extracted from EEG signals. The loop between brain and computer was closed by presenting new recommendations to the users according to the EEG-based feedback, which resulted in a significant information gain for about 70% of the participants of the study. This work constitutes presumably the first proof-of-concept IR systems that have performed automatic information filtering on the basis of brain activity alone.

Despite these advancement such as studies of neurophysiologic correlates of relevance, and applications using different kinds of stimuli, there is a lack of understanding on how to integrate neurophysiology-based relevance feedback in a realistic IR scenario. On one hand this includes the need of standardized approaches and procedures in research (Mostafa & Gwizdka, 2016) considering for example the use of machine learning. More importantly questions arise on what user intent and retrieval models are best suited to process the obtained implicit relevance feedback and how this can be combined with other relevance information obtained for example through explicit feedback.

An Approach for Single-Trial Relevance Computation in IR

Uncertainty in Single-Trial EEG Decoding

Because of the comparably high conductivity of the brain and scalp with respect to the one of the skull, electrical signals arrive spatially smeared at the EEG sensors, leading to low signal-to-noise ratio. Each sensor receives a mixture of signals from many sources in the brain and, conversely, the signals of one particular brain source are recorded at many different electrodes with a broad spatial profile. The predominant approach for real-time decoding is to employ multivariate data analysis methods from the field of machine learning (Lemm, Blankertz, Dickhaus, & Müller, 2011) and to train subject-specific decoding models on calibration data. Although this approach is comparably effective, a high degree of uncertainty in single-trial analysis remains, probably because of the very high number of potentially disturbing sources.

DOI: 10.1002/asi

The perception and cognitive evaluation of visual stimuli, such as information presented on a computer screen, is reflected by event-related potentials (ERPs). In the well-known ERP-based *Row-Column Speller* (Farwell & Donchin, 1988), users concentrate on a target symbol while the rows and columns of the matrix of all symbols are flashing randomly. If the user fixates on the target symbol by gaze, the detection tasks boil down to a mere detection of flashes. More recent ERP-based spellers, such as the *Center Speller* (Treder, Schmidt, & Blankertz, 2011) circumvent the gaze-dependency of the *Row-Column Speller* by posing a higher load on the user as it requires the recognition of a target shape or color. Advancing further into the realm of IR (3), the evaluation of information involves semantic interpretation and more complex stimuli with large variability. In this escalation, the brain responses follow an increasingly less common temporal structure across trials. This leads to a larger variability in the latencies, but also in the morphology of the ERPs and, therefore, to a larger uncertainty in the decoding, see Figure 1.

The challenge of extracting information from a single-trial EEG gets even larger when free-viewing applications are considered. A suitable method for the investigation of free-viewing tasks are eye-fixation-related potentials (EFRP), see (Baccino & Manunta, 2005). Nevertheless, the decoding of the cognitive processes is hampered. On one hand, further unrelated brain activity connected to saccades and artifacts from eye movements overlay the EEG and, on the other hand, the temporal relationship between target-related ERP components and eye movements is variable because task-relevant processing of visual objects may already start before the beginning of a saccade, for example when the visual object is still at a peripheral location (Wenzel, Golenia, & Blankertz, 2016).

Neurophysiology-Based Relevance Computation

We propose a method to predict the relevance of textual keywords from brain signals and eye movements. The approach follows a supervised learning scheme, in which a user-specific classifier is trained by using labeled data. Then, the trained classifier can be used to generate relevance measures online, which can potentially be used in a feedback loop while the user interacts with the system. This machine learning approach is parallel to most modern BCI systems (Nijholt et al., 2008).

Training the classifier. The purpose of this first phase (referred as "the calibration phase") is to gather enough brain activity associated with the user's relevance judgments to train a classifier that will then be used to generate relevance measures online. A series of keywords for which relevance labels are known are presented to the user, and eye tracking is employed to identify when an eye fixation falls on a keyword. For each fixation that falls on a keyword, a high-dimensional feature vector is extracted from the EEG and eye movements (see below) and is labeled as "relevant" or "irrelevant" according to the known label of the keyword. A classification function is then trained to discriminate the feature vectors of the "relevant" and the "irrelevant" classes. To this end, regularized linear discriminant analysis is used (Friedman, 1989), whereby the shrinkage parameter is calculated with an analytic method (Ledoit & Wolf, 2004; Schäfer & Strimmer, 2005).

Online relevance computation. Once the system has been calibrated for the specific user by training a user-specific classifier, the user can interact with the system while EEG signals and eye movements are monitored (referred to as "the online phase"). For each keyword fixated on, a high-dimensional feature vector is extracted (see below), and the classifier infers its label online as belonging to the "relevant" or "irrelevant" classes. This means that the relevance predictions are available to the system in real time and can be used in an adaptive feedback loop.

Feature extraction. High-dimensional feature vectors are extracted from EEG channels recorded at 1000Hz according to the following steps: First, the multichannel EEG signal is re-referenced to the linked mastoids and low-pass filtered (with a second order Chebyshev filter; 42 Hz pass-band, 49 Hz stop-band). The continuous signal is then segmented by extracting the interval from 100 ms to 800 ms after the onset of every eye fixation. Slow fluctuations in the signal are removed by baseline correction (i.e., by

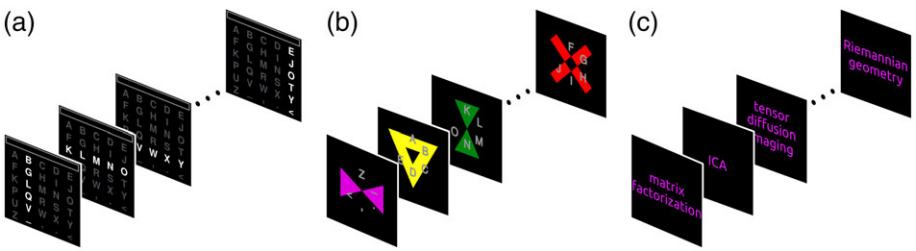

FIG. 1. From target to relevance detection. The classical row-column speller **(a)** which consists essentially in the detection of flashing. The center speller **(b)** relies on the recognition of a target shape/color. In contrast, the task to search for relevant terms **(c)** is incomparably more complex. [Color figure can be viewed at wileyonlinelibrary.com]

DOI: 10.1002/asi

subtracting the mean of the signal within the first 50ms after the fixation onset from each epoch). The signal is downsampled from the original 1000 Hz to 20 Hz to decrease the dimensionality of the feature vectors to be obtained (14 values per channel). A low dimensionality in comparison to the number of available samples has been shown to reduce the risk of overfitting to the training data, which in turn is beneficial for the classification performance (Blankertz, Lemm, Treder, Haufe, & Müller, 2011). The multichannel signal is vectorized by concatenating the values measured at the EEG channels at the 14 time points. The fixation duration is concatenated as an additional feature to the EEG feature vector. Other eye-tracking-related features (e.g., gaze velocity) are not considered as they are not provided in real time by the application programming interface of the device. Further, eye-movement-related signal components are not removed from the EEG because the classifier is expected to deal with task-unrelated eye-movements.

Method validation. To validate the approach in terms of computing relevance measures from semantic words, we carried out a *prior experiment* ($N = 15$). The main question addressed was whether relevance inference from the electroencephalogram (EEG) can be applied in settings where the interpretation of the semantics goes beyond the simple recognition of a previously known letter, picture, or shape that is repeatedly flashed. In the experiment, participants looked for words that belonged to semantic categories, and it was predicted in real-time which words, and thus which semantic category, was the one the user was interested in. Results showed that models using EEG features alone, and in combination with the eye fixation duration feature were able to generate single trial predictions on the keywords significantly above chance levels. Further, these predictions were aggregated in real time to provide reliable estimates of which were the semantic category of interest, showing slight improvements when adding fixation duration to the EEG-based feature vectors. Complete details on the *prior experiment* have been published separately in Wenzel et al. (2017).

The *prior experiment* provided several insights. First, it validated the use of EEG and eye gaze signals to infer subjective relevance of words that required interpretation with respect to their semantics in a free search task (as opposed to commonly used "counting" tasks). Further, predictions were generated on words that were presented simultaneously, relating neural activity to keywords using eye tracking. The *prior experiment* also evidenced the relatively low single-trial classification performances, which were successfully dealt with in real time by averaging over semantic categories. However, when interacting with a real IR system, the user interest and intentions may be more complex than as simulated in the *prior experiment*, and other mechanisms should be envisaged to integrate contextual information that may help to correct the noisy single-trial prediction accuracies.

Addressing Uncertainty in an Online Neuroadaptive System through Interactive Intent Modeling

A promising solution to cope with the uncertainty in the user's intent is interactive intent modeling (Ruotsalo, Jacucci, Myllymäki, & Kaski, 2015), where the potential search intentions of the user are represented and visualized as keywords, their relevance are estimated using feedback signals from the user, and information corresponding to the model is retrieved. In terms of neuroadaptive systems, intent modeling can mitigate both the uncertainty related to the noise present in neurophysiologic signals and the mismatch between the user's articulation of information needs and the encodings of the information to be retrieved.

Adapting the intent model from suboptimal and noisy user feedback

The intent model directly couples the potentially suboptimal user feedback originating from implicit and explicit user signals. The implicit feedback is connected to explicit feedback by considering source-specific probabilistic assumptions on their uncertainties. This provides the flexibility to learn the true uncertainty of each feedback given all preceding feedback.

Estimating the intent model. The relevance of keywords in the model is described with a linear Gaussian model, with which the accuracy of the feedback may differ for the different source types (implicit or explicit). The relevance of keyword i is modeled as

$$y_i \sim N\left(x_i\phi, \sigma^2/w_i\right), \tag{1}$$

where x_i is the feature vector representing that keyword, ϕ is the unknown weight vector which is shared between all keywords and maps the feature vectors to relevance values representing user intent, σ^2 is the variance of feedback noise, and w_i models the accuracy of the relevance feedback. We assume prior distributions on the parameters to be

$$\phi \sim N(0, \lambda I),$$
$$\sigma^2 \sim InverseGamma(\alpha_{\sigma^2}, \beta_{\sigma^2}),$$
$$w_i \sim Gamma(\alpha_w, \beta_w),$$

where λ, α_{σ^2}, and β_{σ^2} are fixed hyperparameters. A key aspect of our approach is that we distinguish between implicit and explicit feedback by using different hyperparameters for prior of the accuracy values, that is, $\left(\alpha_w^{exp}, \beta_w^{exp}\right)$ for explicit feedback and $\left(\alpha_w^{imp}, \beta_w^{imp}\right)$ for implicit feedback.

The posterior of the model estimates both the user's current search intent (ϕ) and the accuracy of the user relevance feedback (w_is). As mentioned, the accuracies of the user feedback on keywords are unknown and drawn from

DOI: 10.1002/asi

a gamma distribution with two parameters: alpha and beta. The model differentiates among explicit and implicit feedback by using different sets of hyper-parameters for the gamma distribution. The explicit feedback is considered very certain (a gamma distribution with mean 1 and very small variance, that is, $\alpha_w^{exp} = 100, \beta_w^{exp} = 100$). On the other hand, the implicit feedback is uncertain *a priori* (gamma distribution with mean 0.5 and large variance, that is, $\alpha_w^{imp} = 1, \beta_w^{imp} = 2$), and therefore, its accuracy is mostly inferred from observations. For example, if the implicit feedback is in line with the previous history of feedback, then it will be inferred as certain and will contribute to the user model. However, if it contradicts the system's current belief, learned from sequence of feedback, then its accuracy may be inferred as a low value and it will not affect the user model (the posterior of ϕ) much. The model infers the true accuracies and corrects the noise in the feedback. We use mean-field variational inference for the posterior inference (Attias, 1999; Kangasrääsiö, Chen, Glowacka, & Kaski, 2016).

Estimating Document Relevance

In addition to estimating the relevances for the keywords in the intent model, the relevances of the documents are estimated and ranked. We employ the feature transformation that projects the relevances estimated for the keywords to the documents (Daee et al., 2016). The underlying principle is that the transformation projects documents in the feature space of the keywords as the relevance of a document is a weighted sum of the relevance of individual keywords that have appeared in it. Based on this projection, the relevance of a document also follows Equation 1 with the difference that the document feature vector is generated from the feature projection.

Exploring uncertainty. Estimating the intent model by directly exploiting the feedback observed from the user yields to showing items like those already judged relevant by the user in the previous iterations. Because the implicit feedback observed from the user may be inaccurate, this exploitative choice might cause the intent model to converge to a suboptimal representation of the user's intention. Alternatively, the system might exploratively select items that are relevant, but also uncertain. These items are likely to be better for obtaining feedback in subsequent iterations as they are novel and not too similar to the ones already judged by the user.

Multiarmed bandits have been shown to be able to model this exploration and exploitation dilemma in information seeking (Ruotsalo et al., 2015). We use the Thompson sampling algorithm (Agrawal & Goyal, 2013) as a solution to the multiarmed bandit problem, to control the exploration and exploitation balance of the recommended keywords and documents (Daee et al., 2016). The idea behind Thompson sampling is that the uncertainty in the marginal posterior of ϕ can by itself control the exploration and exploitation of the items. To implement the algorithm, it is enough to draw a sample from the posterior and rank all the keywords and documents accordingly. In detail, the Thompson sampling algorithm performs the following steps in each iteration:

1. Draw a sample from the marginal posterior of ϕ and denote it as ϕ^p.
2. Rank all the keywords based on the inner product $x_i^T \phi^p$.
3. Rank all the documents based on the inner product $x_j^T \phi^p$.
4. Recommend the highest ranked items and gather the feedback.
5. Update the posterior.

Here, x_i and x_j denote the feature vectors of keyword i and document j (after the transformation) respectively. The highest ranked recommendations were expected to consider the balance between exploration and exploitation (Agrawal & Goyal, 2013).

Visualizing the Intent Model for Explicit and Implicit Interaction

To enable implicit and explicit feedback from the user, the intent model needs to be visualized for interaction. The implicit feedback is captured via capturing eye fixations and EEG signal.

Interface views. The interface consists of two separate views: intent model view and document view. The intent model view, shown in Figure 2, visualizes the top-k keywords chosen based on their estimated weights resulting from the Thompson sampling algorithm. The view employs a circular layout chosen to increase eye tracking accuracy, which is higher at the center of the screen. The keyword are positioned randomly but the layout is optimized to increase the distance between neighboring keywords for more robust matching with eye fixations. The document view, shown in Figure 3, has a conventional ranked list visualization.

Interaction. The search is initiated by entering a query, which results in the first set of results retrieved by the system. To direct the search, users can open a view that displays a set of keywords that are potentially relevant to the users' search intent. The users can examine these keywords and provide explicit relevance feedback on one of the keywords by clicking on it. Although users examine the keywords, the physiologic classifier generates implicit relevance feedback on them. The system then updates the intent model by taking into account both the explicit relevance feedback, and the implicit feedback generated from the keywords the user fixated on. The system then returns the next iteration of results. This process is repeated until the user decides to change the query or ends the search task. Figure 4 depicts the user-system interaction as a control loop.

DOI: 10.1002/asi

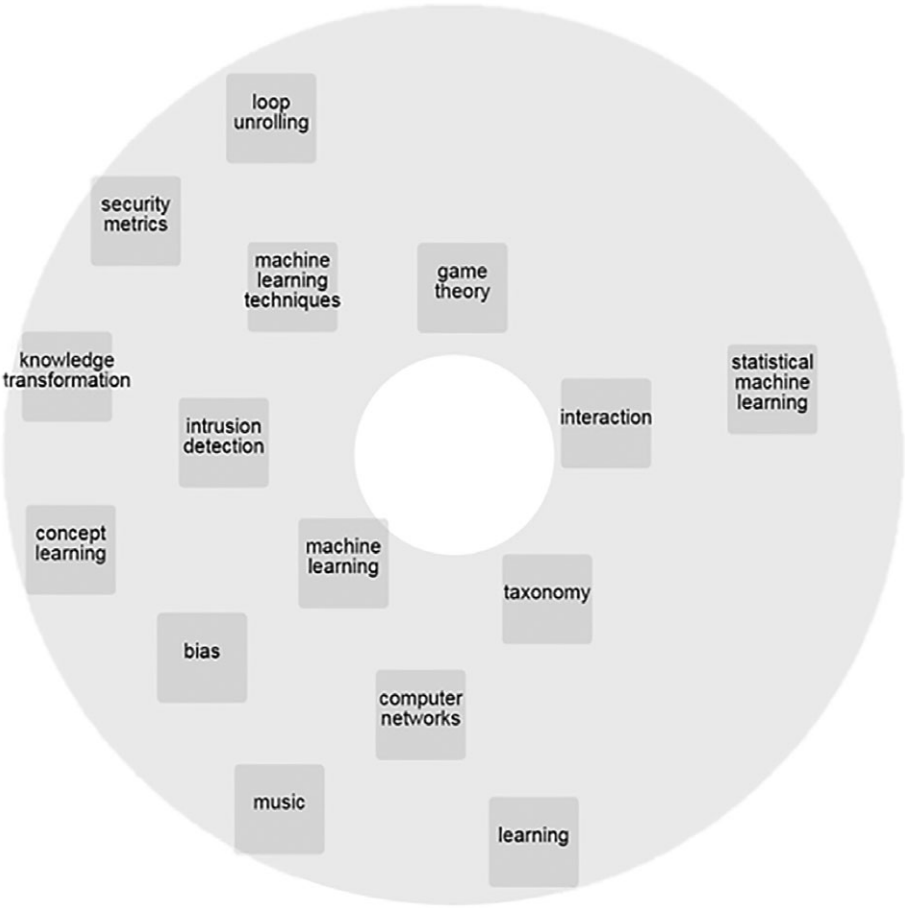

FIG. 2. A screenshot of the user interface displaying the intent model view.

An Experiment in Neuroadaptive Literature Search

This experiment helps to evaluate the approach and system presented in the previous two sections by investigating the following questions:

Is it possible to predict online relevance from neurophysiology in a realistic search task and integrate it as implicit feedback in combination with explicit feedback in interactive intent modeling?

System Apparatus

The system that integrates neurophysiology-based implicit feedback with interactive intent modeling is implemented as a web application using a frontend (the *interface*) - backend (the *engine*) architecture, see Figure 5. The engine comprises of three main components: The *Controller*, which coordinates the different components of the system; the *Physiologic Classifier*, which generates real-time implicit relevance feedback, and the *Interactive Intent Model*, which handles the user model and the information items of the system. The *Physiologic Classifier* is implemented within the framework of the BBCI-Toolbox.[1] For each gaze-fixation, the classifier sends to the *Controller* a relevance value. The *Controller* checks whether the fixation falls on a keyword visible on the screen to associate the predicted relevance value to it. For collecting eye movements, the system uses the SensoMotoric Instruments RED500 eye tracker, interfaced through the SMI iViewX SDK.[2] For collecting brain signals, the system supports the BrainProducts QuickAmp and BrainAmp amplifiers,[3] both of which recorded 32 EEG channels at a sampling rate of 1000 Hz. The *Interactive Intent Model* uses the same document-retrieval model as in Ruotsalo et al. (2013) to select subset of documents, and uses a data set from the following data sources: the Web of Science prepared by Thomson Reuters, Inc., the digital library of the Institute of Electrical and Electronics Engineers (IEEE), the digital library of the Association of Computing Machinery (ACM), and the digital library of Springer. The hyperparameters of the intent model were tuned as $\alpha_{\sigma^2} = 2$, $\beta_{\sigma^2} = 0.1$, and $\lambda = 0.1$ based on pilot experiments (N = 27).

Participants

Sixteen participants (3 females) took part in the experiment. The participants ranged from 22 to 39 years old

[1] https:/github.com/bbci/bbci_public

[2] http://www.smivision.com/

[3] http://www.brainproducts.com/

DOI: 10.1002/asi

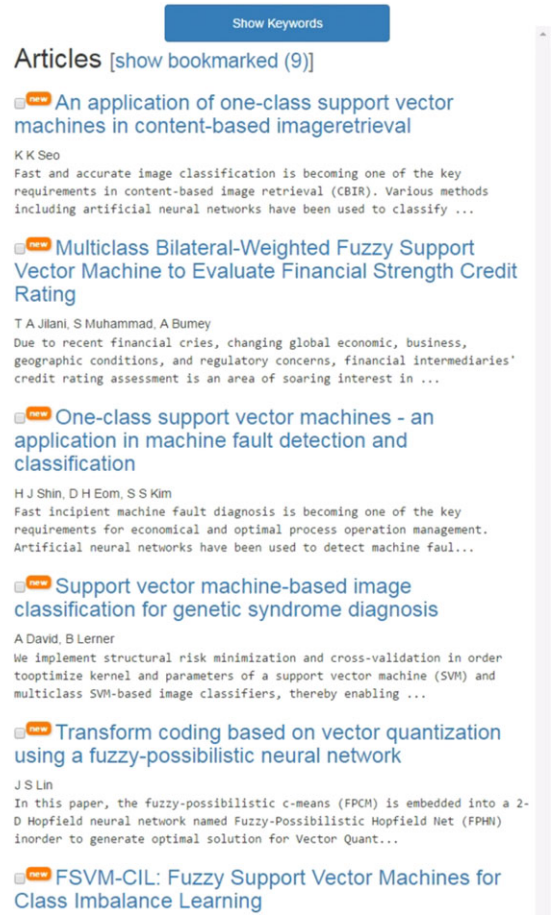

FIG. 3. A screenshot of the user interface displaying the document view. [Color figure can be viewed at wileyonlinelibrary.com]

(M = 28.3). Three participants were postdoctoral researchers, and the rest were students (8 post-graduate, 5 undergraduate) from the University of Helsinki in Finland and the University of Padova in Italy. The participants reported themselves as being physically and mentally healthy. The participants reported a good level of English (M = 4.0, SD = 0.9, on a 1 to 5 scale) and high expertise in computer science (M = 4.4, SD = 0.6, on a 1 to 5 scale). Their experience with browsing scientific literature (M = 3.6, SD = 0.9, on a 1 to 5 scale) and their prior knowledge of machine learning (M = 2.8, SD = 1.5, on a 1 to 5 scale) varied.

Procedure and Experimental Task

At the beginning of the session, the participants were welcomed and briefed as to the procedure and purpose of the experiment before signing the informed consent form. The participants were instructed about the duration of the experiment and reminded that they could withdraw from the experiment at any point in time, without facing negative consequences. Although the physiologic sensors were set up, the participants filled a background information questionnaire. Following, a standard 9-point eye tracker calibration procedure was carried out repeatedly until reaching an error smaller than 0.5 degrees of visual angle.

The calibration phase. The participants then engaged in the calibration phase for around 1 hour, until the system had collected enough data points to train the physiologic classifier. The participants were allowed to have small breaks during the calibration phase whenever they felt tired or their concentration was diminishing. To collect training data for the physiologic classifier, we generated a data set that matched the application domain by using a subset of the data set used by the interactive intent model system. The data set consisted of a set of topics with associated keywords and was created using expert judgments in an iterative process that aimed at minimizing the overlaps between the topics, while maximizing the dissociation between relevant and irrelevant keywords to a given topic.[4]

Participants were prompted with a list of five topics, randomly selected from the calibration data set. On selecting a topic, a series of keywords were shown to the user, who was asked to select the keywords relevant to the topic. This procedure was repeated iteratively for several topics, until the system had gathered enough data to train the physiologic classifier.[5]

The online phase. Once enough data had been collected and the physiologic classifier had been trained, the participants engaged in the online phase. Participants were provided the following instructions:

Imagine that you are going to write an essay about topic X. Please bookmark the articles on the scroll list that you think are relevant to the topic, so that you can use them later in the essay. You will later be asked to write a short outline of the essay based on your bookmarked articles.

The participants had to perform two versions of the same task, using the topics "neural networks" and "support vector machines." One of the tasks was performed using the full system. The other task was performed using a baseline system, which behaved in the exact same way as the full system, but no implicit relevance feedback was fed to the interactive intent model system. Instead, only the explicit feedback provided by the user was used to refine the user model and present the next iteration of results. The participants were unaware that they were using two different systems, and they were naïve about the systems' implementation.

For evaluation purposes, the participants were prompted at the end of each iteration with a dialog asking them to label the relevance of the keywords they had fixated on (on a scale from 0 to 5). This allowed the "ground truth" to be collected on the relevance of the presented keywords as perceived by the users. This was otherwise not

[4] For review: Refer to *Appendix A* for more details on the generation of the calibration data set.

[5] For review: Refer to *Appendix B* for details on how the assessment of keywords' relevance was carried out by the participants during the calibration phase.

DOI: 10.1002/asi

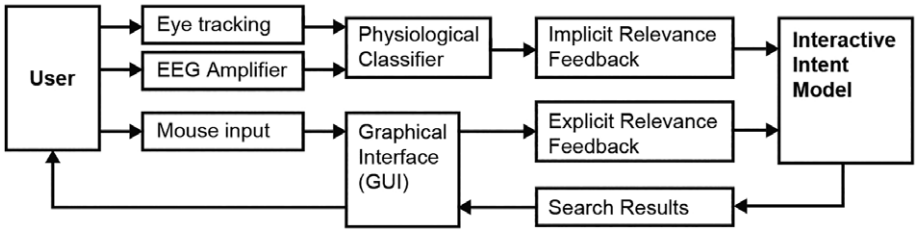

FIG. 4. Summary of the system as a control loop during the online phase.

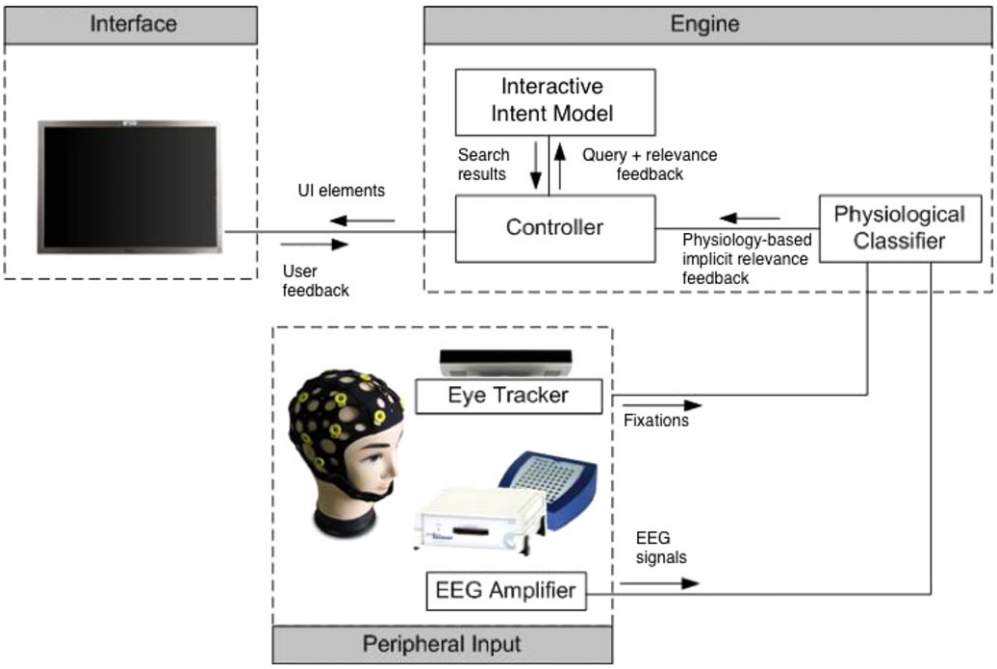

FIG. 5. Components of the system. [Color figure can be viewed at wileyonlinelibrary.com]

available, as the keywords were generated in real-time from the interactive intent model system, and their relevance naturally depends on the users' information needs, which were not known *a priori.*

The participants performed each task in the online phase for around 20 minutes, for a maximum of 10 iterations. The task and system type were counterbalanced. On completion of the task, participants were rewarded with two movie tickets. In total, the experiment lasted approximately 2.5 hours.

Measures and Analyses

Calibration phase. To evaluate the feasibility and performance of the system in predicting relevance from brain signals, we first evaluated the classification performance in the calibration phase. The data used in the calibration phase were controlled and had the advantage that the same data set was used to train the different user-specific classification models. Classification performance was computed in terms of area under the ROC curve (AUROC) and was evaluated using a standard 10×10 fold cross validation approach. AUROC is a widely used and sensible measure, even under class imbalances, that links the *true positive rate* and the *false positive rate* while avoiding possible misinterpretations such as the accuracy paradox (Zhu & Davidson, 2007).

To quantify the significance and the effect sizes of the implicit relevance feedback from the brain signals, we compared the classification performances against performances from prediction models learned from randomized labels. Standard permutation tests were applied for significance testing (Good, 2013). In detail, for each of the 16 participants, we ran within-participant permutation tests with 1000 iterations. For each iteration, we learned a classification model using randomized labels, and we then computed the p-value as the percentage of random classification performances that were equal to or greater than the true classification performance.

Online phase. The aim was to assess how well the classification performance achieved in the calibration phase transferred to the online phase, during which the users were engaged in a realistic information-seeking task, and

DOI: 10.1002/asi

the data presented to the user from which implicit relevance feedback was classified were generated in real-time. To do so, for each participant we computed the classification performance in terms of AUROC for each of the fixated keywords in the online phase in the tasks for which the participants used the full system. We used the feedback provided by the participants on the keywords as the labels. We binarized the user feedback, so that keywords that were rated between 0 and 2 were considered irrelevant and keywords that were rated between 3 and 5 were considered relevant. Further, participants *P05* and *P06* had to be rejected from the analysis because the server hosting the interactive intent model system went down during the execution of the online phase.

As explained in Section *Addressing Uncertainty in an Online Neuroadaptive System through Interactive Intent Modeling*, in each iteration, the intent model learns the relevance of all keywords from the available sequence of explicit and implicit feedback. Accordingly, we also computed the classification performance in terms of the AUROC of the relevance of keywords estimated by the intent model. This is the performance after the user model has accounted for the noise in implicit relevance feedback values coming from the physiologic classifier.

Task performance. After completion of the search task, participants were asked to write down some of the concepts that they had learned about the topics, which lead to a very heterogeneous collection of "mini-essays" not suited for comparison across participants. Instead, to assess whether using physiology-based implicit relevance measures had an influence on the task performance, we compared the quality of the documents that participants bookmarked when using the full system (including implicit relevance feedback) and when using the baseline system (that did not include implicit relevance feedback). In total, participants generated 397 bookmarks on 277 different documents. On the population level, documents had often been bookmarked using both system types (e.g, one participant had bookmarked a document in the baseline system condition, while another participant had bookmarked the same document in the full system condition). To assess any change in task performance between the two conditions, we therefore limited the analysis to a "representative" subset of documents that minimized overlaps. Documents were selected as "representative" for one of the conditions if on the population level, the document was bookmarked two or more times than in the other condition. This lead to a subset of 21 documents (8 representing the full system condition, and 13 representing the baseline system condition). Documents were then rated by 3 experts (on a 1-6 rating scale), on their *relevance* (i.e., is this document relevant to the search task), *obviousness* (i.e., is this a well-known overview article in a given research area), and *novelty* (i.e., is this article uncommon yet relevant to a given topic or specific subtopic in a given research area) (Ruotsalo et al., 2013). Ratings were averaged across experts, and Wilcoxon rank-sum tests were used to test for statistical differences between the two conditions (full system vs. baseline system), for each of the three rating categories (*relevance*, *obviousness*, and *novelty*).

Results

Calibration phase. Classification performance proved to be significantly better than random for 13 out of 16 participants, meaning that we were able to successfully train the classifier for around 80% of the participants. On the population level, AUROC resulted in 0.61 $\pm$ 0.02 (mean $\pm$ standard error of the mean). Figure 6 presents the individual classification performances in the calibration phase.

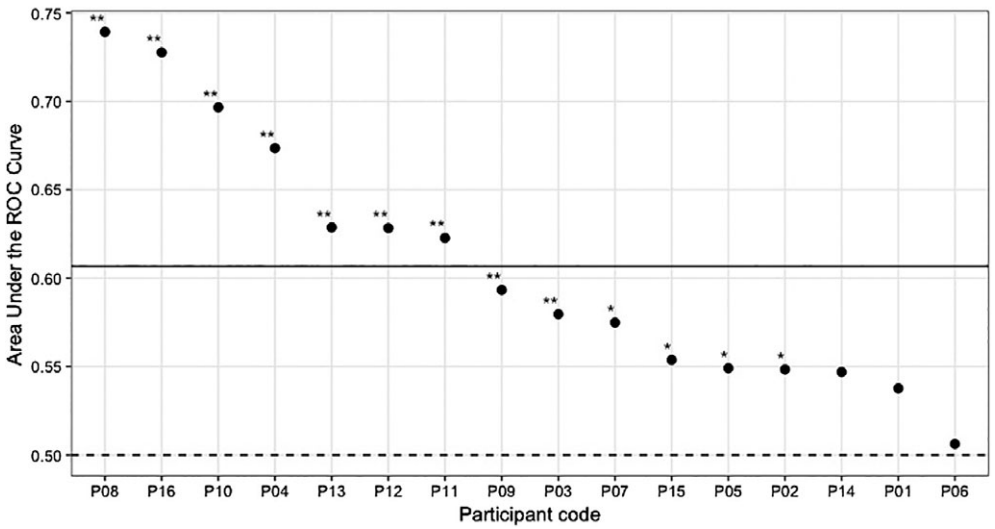

FIG. 6. Individual classification performances in the calibration phase in terms of area under the ROC curve (AUROC), and improvement over the random baseline at the levels of $p < 0.05$ (*), and $p < 0.001$ (**). The horizontal lines represent the mean (solid) and random (dashed).

DOI: 10.1002/asi

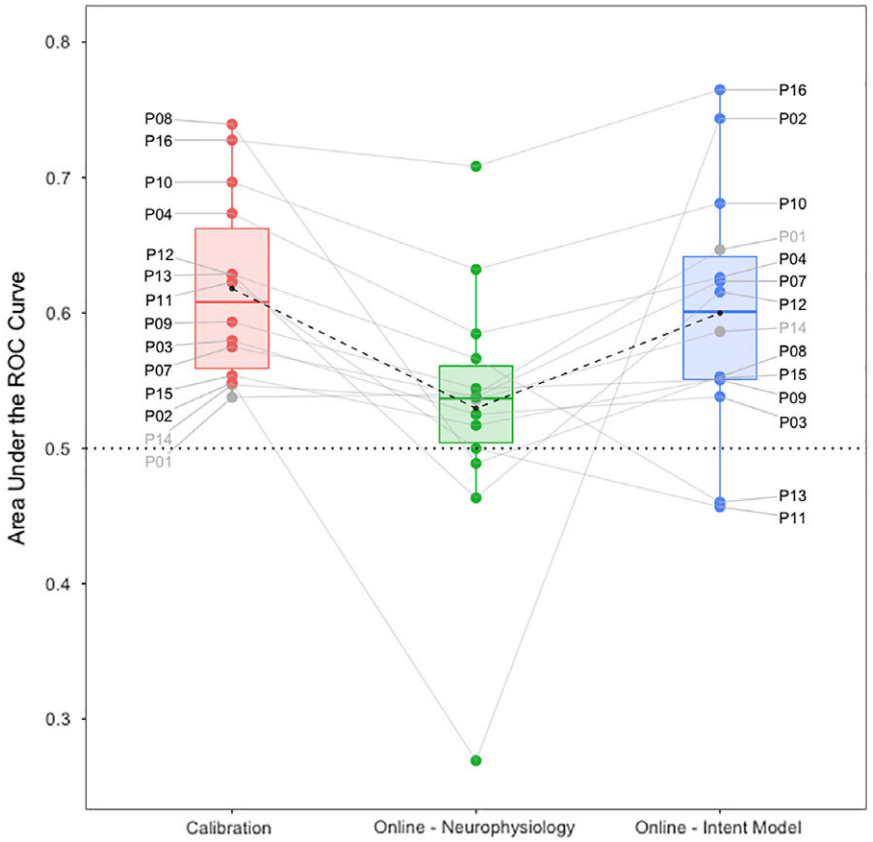

FIG. 7. Individual classification performance in terms of area under the ROC curve (AUROC). Left: offline prediction in the "calibration phase." Middle: neurophysiologic prediction in the "online phase." Right: intent model prediction in the "online phase." Smaller black dots and dashed lines indicate mean classification performance. The dashed horizontal line represents random classification. Participants for which calibration did not outperform random predictions are presented in gray. [Color figure can be viewed at wileyonlinelibrary.com]

Online phase. Online relevance predictions as directly obtained through the physiologic classifier presented averaged AUROC values on the population level of 0.53 ± 0.03 (mean ± standard error of the mean). The performance was improved by the user model, leading to averaged AUROC values of 0.60 ± 0.03. In fact, the intent model increased prediction performance for 10 out of the 12 participants for which a classifier was successfully trained in the calibration phase, representing over 80% of these participants. Figure 7 shows the results of the classification performance for the calibration phase and for the online phase, in terms of the implicit relevance feedback, both as directly obtained through classification of brain signals, and as inferred by the intent model.

Task performance. Wilcoxon rank-sum tests did not show statistical difference between the full system and baseline system, for any of the rating categories: In terms of *relevance*, expert ratings provided to representative documents of the full system (*Mdn* = 3.5) did not significantly differ from those of the baseline system (*Mdn* = 4.67), $W = 69$, $p = 0.22$. In terms of *obviousness*, expert ratings provided to representative documents of the full system (*Mdn* = 2.67) did not significantly differ from those of the baseline system (*Mdn* = 3.33), $W = 73.5$, $p = 0.12$. In terms of *novelty*, expert ratings provided to representative documents of the full system (*Mdn* = 3.83) did not significantly differ from those of the baseline system (*Mdn* = 3.67), $W = 55.5$, $p = 0.82$.

Discussion and Conclusions

A methodology for predicting implicit relevance feedback from human-brain measurements in a realistic search scenario was presented. The methodology was implemented in a first-of-its-kind interactive IR system that combines brain-based feedback and eye tracking with scarce explicit feedback. To our knowledge, the presented system is the first closed-loop IR system that utilizes brain-based feedback, combines it with eye-tracking and explicit feedback, and is evaluated in realistic IR tasks.

Empiric Findings

The empiric evidence suggests that the presented methodology allows to reliably train classification models for implicit relevance prediction by using complex real-world data. The results show that the classification performance significantly outperforms random predictions for over 80%

DOI: 10.1002/asi

of the participants, with some of the participants reaching AUROC values over 0.7. One explanation for the random classification outcomes among the remaining approximately 20% of participants could be the fact that BCI control does not work for a non-negligible proportion of users (approximately 15 - 30%) (Acqualagna, Botrel, Vidaurre, Kübler, & Blankertz, 2016; Allison et al., 2010; Blankertz et al., 2010; Guger et al., 2009). These results are comparable to the ones obtained in the *prior experiment* (see Section *Validating the Relevance Computation Method*, and (Wenzel et al., 2017)), where a limited and controlled data set of keywords was used.

In addition, the results show that the classification performances achieved using the controlled "calibration data set" in the calibration phase transferred to the online phase, during which the retrieved documents and keyword varied for each participant, and their perception of relevance was related to their current information needs, rather than to a predefined experimental task. Although the classification performance decreased as expected, the overall distribution across participants remained above random classification levels.

Further, we demonstrate that the approach is able to combine the noisy neurophysiology-based implicit relevance feedback with limited explicit feedback (one per search iteration), which improved the classification performance for over 80% of the participants for which we had successfully trained a classifier during the calibration phase.

Figure 7 shows atypical values for participant *P02*. By looking at the data, we found out that this participant provided highly unbalanced ground truth in the online phase (i.e., 96% of the ground truth provided was from the relevant class), which explains the drastic changes in the AUROC values. Thus, the magnitude of such changes in the performance measures should be interpreted cautiously. Further, we looked at the participants who consistently presented high AUROC values (i.e., *P04*, *P10*, and *P16*) to identify factors that could explain their better performance, which could potentially be used to improve the overall model and the design of future studies (e.g., level of education, prior knowledge about the topic, reported satisfaction and engagement with the system, etc.). We did not find anything especially noteworthy about them, nor performance differences between undergraduate and postgraduate participants.

Limitations

Our approach and study includes at least the following limitations. First, the predicted relevance from physiology, although promising, still leaves room for improvement, both in terms of classification performance and uniformity across participants. Second, as the online phase involves an online interaction loop between the participants and the system, we could not perform offline permutation tests to evaluate how much did the achieved classification performances improved over random classifications (as done for the calibration phase). This would have provided further empiric insight on how much the classification performance transferred from the controlled calibration phase to the realistic online phase, as well as on how much of the performance of the intent model is explained by the implicit feedback and how much by the scarce explicit feedback. Third, the analysis on the selection behavior of bookmarked documents did not yet yield conclusive results in terms of task performance improvements. Future work should extend the presented results by further studying how the reported classification performances could transfer over to search task performance. Overall, although we report on the first-of-its-kind closed-loop information retrieval system that fully integrates neurophysiologic signals while users perform real information seeking task, the experimental method and results indicate that there is still room for improvement, both to demonstrate the impact of the implicit feedback to the overall intent model, as well as to exemplify how using neurophysiologic input transfers to improved task performance.

Implications and Future Work

In essence, relevance judgments happen in the human brain and therefore the most intriguing way to predict relevance is to directly utilize the brain signals. These signals have advantages over the more conventional sources of user signals from a practical IR point of view. The recording of the relevance judgments directly from human neurophysiology do not require any explicit user interaction, such as user actively clicking on items. The current work contributes showing that predicting the relevance from neurophysiology on information presented in realistic information retrieval system responses is possible with promising accuracy. Moreover, it is demonstrated that the relevance prediction methodology can be operationalized as a part of a closed-loop information retrieval system. In concrete the work contributes not only a reference approach and procedure but proposes interactive intent modeling (Ruotsalo et al., 2015) as a promising user intent and retrieval model suited for processing implicit relevance and combining it with other relevance information such as explicit feedback. Our findings open a horizon for information retrieval systems that can detect relevance directly from human neurophysiology and combine this with potentially scarce explicit signals without requiring users to devote attention for laborious explicit interaction. Future studies can put more emphasis on demonstrating the actual task performance improvement that can be obtained, and can devise ways to collect repeated measures of implicit responses to reduce uncertainty either from the same participant or across participants. The inherent uncertainty in both the brain measurements, attentional focus of the user, and in data representations, however, call for computational methods for simultaneously modeling cognitional states and the data for which these states are associated with. Our findings already show a path towards closed-

DOI: 10.1002/asi

loop systems that are able to analyze and utilize relevance and human cognition directly from wearable sensors as it is manifested as a part of human information search activities.

Acknowledgments

We thank Mats Sjöberg, Antti Kangasrääsiö, Nishadh Aluthge, and Hassan Abbas for their hard work in implementing the system and running experimental studies. This work has been supported by the European Commision (MindSee FP7-ICT; Grant Agreement #611570).

References

Acqualagna, L., Botrel, L., Vidaurre, C., Kübler, A., & Blankertz, B. (2016). Large-scale assessment of a fully automatic co-adaptive motor imagery-based brain computer interface. PLoS One, 11(2), e0148886. https://doi.org/10.1371/journal.pone.0148886

Agrawal, S., & Goyal, N. (2013). Thompson Sampling for Contextual Bandits with Linear Payoffs. In Proceedings of the 30th International Conference on Machine Learning (pp. 127–135). PMLR.

Allegretti, M., Moshfeghi, Y., Hadjigeorgieva, M., Pollick, F. E., Jose, J. M., & Pasi, G. (2015). When relevance judgement is happening?: An EEG-based study. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 719–722).

Allison, B., Luth, T., Valbuena, D., Teymourian, A., Volosyak, I., & Graser, A. (2010). BCI demographics: How many (and what kinds of) people can use an SSVEP BCI? IEEE Transactions on Neural Systems and Rehabilitation Engineering, 18(2), 107–116.

Arapakis, I., Athanasakos, K., & Jose, J.M. (2010). 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 (pp. 371–378). New York, NY: ACM. https://doi.org/10.1145/1835449.1835512

Arapakis, I., Konstas, I., & Jose, J.M. (2009). Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In Proceedings of the 17th ACM International Conference on Multimedia (pp. 461–470). New York, NY: ACM. https://doi.org/10.1145/1631272.1631336

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational bayes. In Proceedings of the fifteenth conference on uncertainty in artificial intelligence (pp. 21–30). San Francisco, CA: Morgan Kaufmann Publishers Inc.

Baccino, T., & Manunta, Y. (2005). Eye-fixation-related potentials: Insight into parafoveal processing. Journal of Psychophysiology, 19(3), 204–215.

Barral, O., Eugster, M.J., Ruotsalo, T., Spapé, M.M., Kosunen, I., Ravaja, N., … Jacucci, G. (2015). Exploring peripheral physiology as a predictor of perceived relevance in information retrieval. In Proceedings of the 20th International Conference on Intelligent User Interfaces (pp. 389–399). New York, NY: ACM. https://doi.org/10.1145/2678025.2701389

Barral, O., Kosunen, I., & Jacucci, G. (2017). No need to laugh out loud: Predicting humor appraisal of comic strips based on physiological signals in a realistic environment. ACM Transactions on Computer-Human Interaction, 24(6), 40. Retrieved from http://doi.acm.org/10.1145/3157730. https://doi.org/10.1145/3157730

Barral, O., Kosunen, I., Ruotsalo, T., Spapé, M.M., Eugster, M.J.A., Ravaja, N., … Jacucci, G. (2016). Extracting relevance and affect information from physiological text annotation. User Modeling and User-Adapted Interaction, 26(5), 493–520. https://doi.org/10.1007/s11257-016-9184-8

Blankertz, B., Lemm, S., Treder, M., Haufe, S., & Müller, K.-R. (2011). Single-trial analysis and classification of ERP components – A tutorial. NeuroImage, 56(2), 814–825. https://doi.org/10.1016/j.neuroimage.2010.06.048

Blankertz, B., Sannelli, C., Halder, S., Hammer, E.M., Kübler, A., Müller, K.-R., … Dickhaus, T. (2010). Neurophysiological predictor of SMR-based BCI performance. NeuroImage, 51(4), 1303–1309. https://doi.org/10.1016/j.neuroimage.2010.03.022

Cowley, B., Filetti, M., Lukander, K., Torniainen, J., Henelius, A., Ahonen, L., … Jacucci, G. (2016). The psychophysiology primer: A guide to methods and a broad review with a focus on human–computer interaction. Foundations and Trends[textregistered] in Human—Computer Interaction, 9(3-4), 151–308. Retrieved from https://doi.org/10.1561/1100000065. https://doi.org/10.1561/1100000065

Daee, P., Pyykkö, J., Glowacka, D., & Kaski, S. (2016). Interactive intent modeling from multiple feedback domains. In Proceedings of the 21st International Conference on Intelligent User Interfaces (pp. 71–75). New York, NY: ACM. https://doi.org/10.1145/2856767.2856803

Eugster, M.J.A., Ruotsalo, T., Spapé, M.M., Barral, O., Ravaja, N., Jacucci, G., & Kaski, S. (2016). Natural brain-information interfaces: Recommending information by relevance inferred from human brain signals. Scientific Reports, 6, 38580. https://doi.org/10.1038/srep38580

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 the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (pp. 425–434). New York, NY: ACM. https://doi.org/10.1145/2600428.2609594

Farwell, L.A., & Donchin, E. (1988). Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. Electroencephalography and clinical Neurophysiology, 70(6), 510–523.

Friedman, J.H. (1989). Regularized discriminant analysis. Journal of the American Statistical Association, 84(405), 165–175. https://doi.org/10.1080/01621459.1989.10478752

Golenia, J.-E., Wenzel, M.A., & Blankertz, B. (2015). Live demonstrator of EEG and eye-tracking input for disambiguation of image search results. In Symbiotic interaction (pp. 81–86). Cham: Springer. https://doi.org/10.1007/978-3-319-24917-9_8

Golenia, J.E., Wenzel, M.A., Bogojeski, M., & Blankertz, B. (2018). Implicit relevance feedback from electroencephalography and eye tracking in image search. Journal of Neural Engineering, 15(2), 026002. https://doi.org/10.1088/1741-2552/aa9999

Good, P. (2013). Permutation tests: a practical guide to resampling methods for testing hypotheses. Berlin/Heidelberg, Germany: Springer Science & Business Media.

Guger, C., Daban, S., Sellers, E., Holzner, C., Krausz, G., Carabalona, R., … Edlinger, G. (2009). How many people are able to control a p300-based brain–computer interface (BCI)? Neuroscience Letters, 462(1), 94–98.

Gwizdka, J. (2014). Characterizing relevance with eye-tracking measures. In Proceedings of the 5th Information Interaction in Context Symposium (pp. 58–67). New York, NY: ACM. https://doi.org/10.1145/2637002.2637011

Gwizdka, J., & Mostafa, J. (2015). NeuroIR 2015: SIGIR 2015 workshop on neuro-physiological methods in IR research. In ACM SIGIR Forum (Vol. 49, pp. 83–88). ACM. https://doi.org/10.1145/2888422.2888435

Gwizdka, J., & Mostafa, J. (2017). NeuroIIR: Challenges in bringing neuroscience to research in human-information interaction. In Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval - CHIIR '17 (pp. 437–438). New York, NY: ACM Press. https://doi.org/10.1145/3020165.3022165

Jacucci, G., Fairclough, S., & Solovey, E.T. (2015). Physiological computing. Computer, 48(10), 12–16. Retrieved from http://ieeexplore.ieee.org/document/7310960/. https://doi.org/10.1109/MC.2015.291

Kangasrääsiö, A., Chen, Y., Glowacka, D., & Kaski, S. (2016). Interactive modeling of concept drift and errors in relevance feedback. In Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization (pp. 185–193). New York, NY: ACM. https://doi.org/10.1145/2930238.2930243

Kauppi, J.-P., Kandemir, M., Saarinen, V.-M., Hirvenkari, L., Parkkonen, L., Klami, A., … Kaski, S. (2015). Towards brain-activity-controlled information retrieval: Decoding image relevance from MEG signals. NeuroImage, 112, 288–298. https://doi.org/10.1016/j.neuroimage.2014.12.079

Kelly, D., & Fu, X. (2006). 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 (pp. 453–460). New York, NY: ACM. https://doi.org/10.1145/1148170.1148249

Kelly, D., & Teevan, J. (2003). Implicit feedback for inferring user preference: A bibliography. SIGIR Forum, 37(2), 18–28. https://doi.org/10.1145/959258.959260

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis, 88(2), 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4

Lemm, S., Blankertz, B., Dickhaus, T., & Müller, K.-R. (2011). Introduction to machine learning for brain imaging. Neuroimage, 56(2), 387–399.

Moshfeghi, Y., & Jose, J.M. (2013). An effective implicit relevance feedback technique using affective, physiological and behavioural features. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 133–142). New York, NY: ACM. https://doi.org/10.1145/2484028.2484074

Moshfeghi, Y., Pinto, L.R., Pollick, F.E., & Jose, J.M. (2013). Understanding relevance: An fMRI study. In European Conference on Information Retrieval (pp. 14–25). Berlin, Heidelberg: Springer.

Mostafa, J., & Gwizdka, J. (2016). Deepening the role of the user: Neurophysiological evidence as a basis for studying and improving search. In Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (pp. 63–70). New York, USA; ACM.

Nijholt, A., Tan, D., Pfurtscheller, G., Brunner, C., Millán, J.d.R., Allison, B., … Müller, K.R. (2008). Brain-computer interfacing for intelligent systems. IEEE Intelligent Systems, 23(3), 72–79. https://doi.org/10.1109/MIS.2008.41

Oliveira, F.T., Aula, A., & Russell, D.M. (2009). Discriminating the relevance of web search results with measures of pupil size. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2209–2212). New York, NY: ACM. https://doi.org/10.1145/1518701.1519038

Puolamäki, K., Salojärvi, J., Savia, E., Simola, J., & Kaski, S. (2005). Combining eye movements and collaborative filtering for proactive information retrieval. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 146–153). New York, USA; ACM.

Ruotsalo, T., Jacucci, G., Myllymäki, P., & Kaski, S. (2015). Interactive intent modeling: Information discovery beyond search. Communications of the ACM, 58(1), 86–92.

Ruotsalo, T., Peltonen, J., Eugster, M., Głowacka, D., Konyushkova, K., Athukorala, K., et al. (2013). Directing exploratory search with interactive intent modeling. In Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management (pp. 1759–1764). New York, USA; ACM.

Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology, 4(1). Article 32. https://doi.org/10.2202/1544-6115.1175

Treder, M.S., Schmidt, N.M., & Blankertz, B. (2011). Gaze-independent brain–computer interfaces based on covert attention and feature attention. Journal of Neural Engineering, 8(6), 066003.

Wenzel, M.A., Bogojeski, M., & Blankertz, B. (2017). Real-time inference of word relevance from electroencephalogram and eye gaze. Journal of Neural Engineering, 14(5), 056007. https://doi.org/10.1088/1741-2552/aa7590

Wenzel, M.A., Golenia, J.-E., & Blankertz, B. (2016). Classification of eye fixation related potentials for variable stimulus saliency. Frontiers in Neuroscience, 10, 23. Retrieved from https://www.frontiersin.org/article/10.3389/fnins.2016.00023. https://doi.org/10.3389/fnins.2016.00023

Zhu, X., & Davidson, I. (2007). Knowledge discovery and data mining: Challenges and realities. Hershey, PA: IGI Global.