Integrating Neurophysiological Relevance Feedback in Intent Modeling for Information Retrieval

Giulio Jacucci[†*], Oswald Barral[†], Pedram Daee[+], Markus Wenzel[§], Baris Serim[†], Tuukka Ruotsalo[†], Patrik Pluchino[‡], Jonathan Freeman[°], Luciano Gamberini [‡], Samuel Kaski[+], Benjamin Blankertz[§]

[†]Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, P.O. Box 68 (Gustaf Hällströmin katu 2b), FI-00014 tel:+358 2941 911 name.surname@helsinki.fi

[+]Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, P.O.Box 15400, FI-00076 Aalto, Finland, tel:+358 9 47001 firstname.lastname@aalto.fi

[§]Neurotechnology Group, Technische Universität Berlin, 10587 Berlin, Germany, tel:+49 30 3140, benjamin.blankertz@tu-berlin.de, markus.wenzel@hhi.fraunhofer.de

[°]Goldsmiths, University of London, New Cross, London, SE14 6NW, UK, tel: +44 20 7919 7884 fax: +44 20 7919 7873, J.Freeman@gold.ac.uk

[‡]Human Inspired Technology Research Centre, University of Padova Via Luzzatti, 4 - 35121 Padova Tel: +39 049 827 5796, name.surname@unipd.it

[*]Corresponding Author: Giulio Jacucci, Department of Computer Science University of Helsinki giulio.jacucci@helsinki.fi +358 2941 51153

Abstract

The use of implicit relevance feedback from neurophysiology could deliver effortless information retrieval. However, both computing neurophysiological responses and retrieving documents are characterized by uncertainty due to 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. While 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).

Keywords: information retrieval, brain-computer interfaces, neuro-physiology, interactive intent modeling, relevance feedback

Integrating Neurophysiological Relevance Feedback in Intent Modeling for Information Retrieval

Introduction

Gathering relevance feedback on information items without disrupting the user is a central challenge in information retrieval (IR). Neurophysiological 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). However, successful application of neurophysiological measures in IR encounters a dual uncertainty problem: (i) noisiness and unknown causes of responses in neurophysiological signals make it difficult to interpret them, a problem exacerbated by the lack of stimulus control in realistic settings, and (ii) 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 click-through activity or brain responses in order 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 utility 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., 2016, 2014). In contrast, realistic search interfaces are characterized by dense information, potential ambiguity regarding the relevance of search results, and user interaction.

After briefly discussing related work on implicit relevance feedback in IR using brain–computer interfaces (BCIs), 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 neurophysiological 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 neurophysiological response with modeling interactively intent in IR. We develop computational techniques that, within an intent model, are able to combine uncertain implicit responses and scarce explicit feedback with intelligent inferences from underlying information modeling. The section presents the first-of-its-kind, fully integrated IR system that makes use of implicit relevance feedback with online computation from brain activity and eye tracking. 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 are able to predict neurophysiology-based relevance feedback in complex data domains and realistic search tasks and combine it with explicit relevance feedback in interactive intent modeling. Our work provides the following contributions:

1. We demonstrate an approach able to predict implicit relevance feedback from neurophysiological measurements in a realistic search scenario.

2. We present a novel interactive IR system that combines in interactive intent modeling noisy brain-based implicit feedback with scarce explicit feedback for better relevance predictions.

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 in order to 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 in order 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), physiological 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., 2016, 2014; Golenia, Wenzel, & Blankertz, 2015; Kauppi et al., 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). Research at the intersection between brain-computer interfaces (BCIs) and IR is still in an early stage, and appropriate neurophysiological 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 in order 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 dataset. Other studies (Golenia et al., 2015; Golenia, Wenzel, Bogojeski, & Blankertz, 2017) demonstrated how the brain response to relevant versus irrelevant information can be harnessed to improve image searches in ambiguous search tasks. 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 there is a lack of understanding on how to integrate neurophysiology based relevance feedback in a realistic IR scenario along with the need of standardized tasks and procedures in research (Mostafa & Gwizdka, 2016).

An Approach for Single-Trial Relevance Computation in IR

Uncertainty in Single-Trial EEG Decoding

Due to 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. While this approach is comparably effective, a high degree of uncertainty in single-trial analysis remains, probably due to the very high number of potentially disturbing sources.

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, as a consequence, to a larger uncertainty in the decoding, see Figure 1.

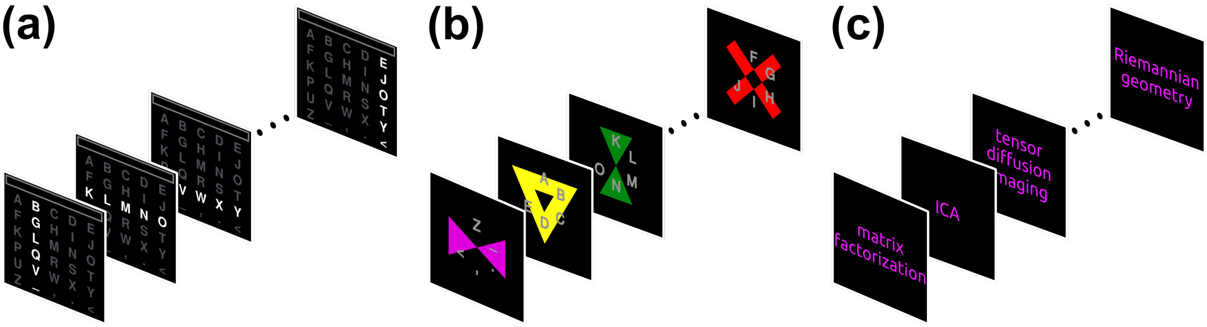

Figure 1. **From target to relevance detection.** The classical row-column speller **(a)** which consists essentially in the detection of flashing. The center speller **(b)** relies on the recognition of a target shape/color. In contrast, the task to search for relevant terms **(c)** is incomparably more complex.

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 since 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 in order 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 upon, 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 multi-channel 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 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 in order 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 multi-channel 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 since the classifier is expected to deal with task-unrelated eye-movements.

Method Validation. In order 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 neurophysiological 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

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

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

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

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, i.e., $(\alpha_w^{exp}, \beta_w^{exp})$ for explicit feedback and $(\alpha_w^{imp}, \beta_w^{imp})$ 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 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, i.e., $\alpha_w^{exp} = 100, \beta_w^{exp} = 100$). On the other hand, the implicit feedback is uncertain *a priori* (gamma distribution with mean 0.5 and large variance, i.e., $\alpha_w^{imp} = 1, \beta_w^{imp} = 2$), 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 similar to those already judged relevant by the user in the previous iterations. Since 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.

Multi-armed 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 multi-armed 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.

In order 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

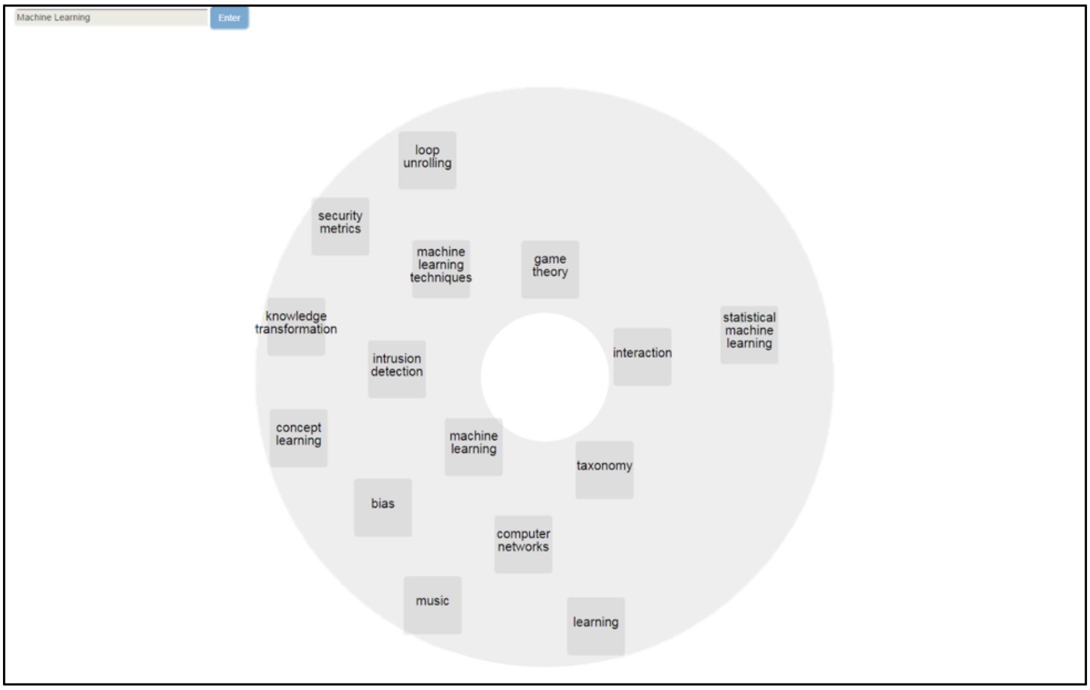

Figure 2. A screenshot of the user interface displaying the intent model view.

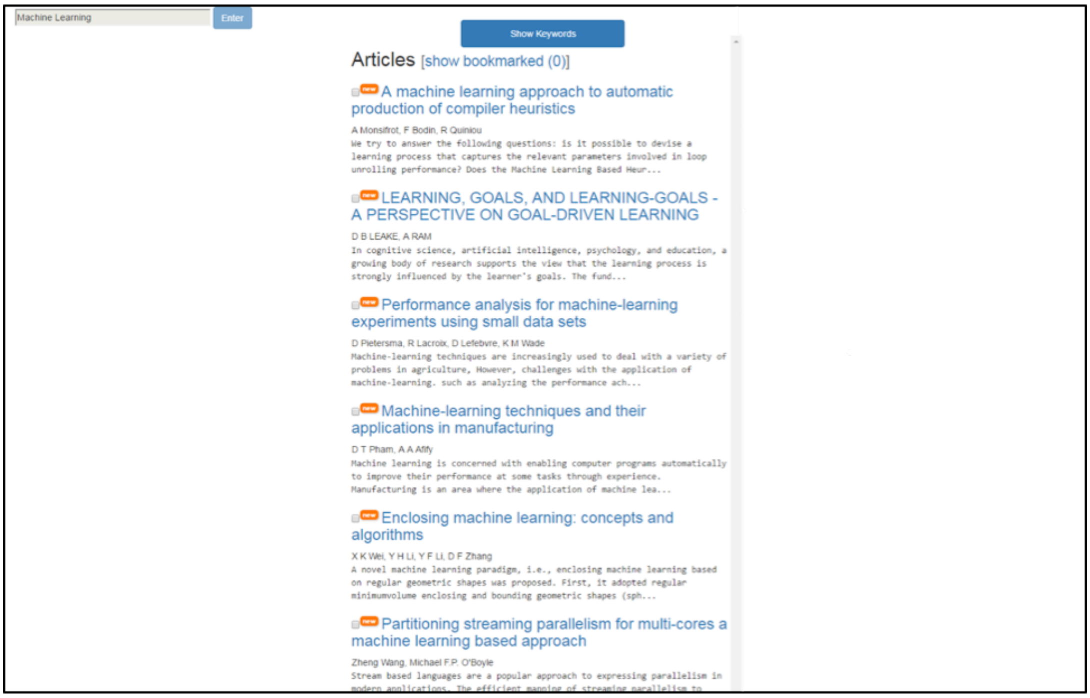

Figure 3. A screenshot of the user interface displaying the document view.

keywords by clicking on it. While users examine the keywords, the physiological 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.

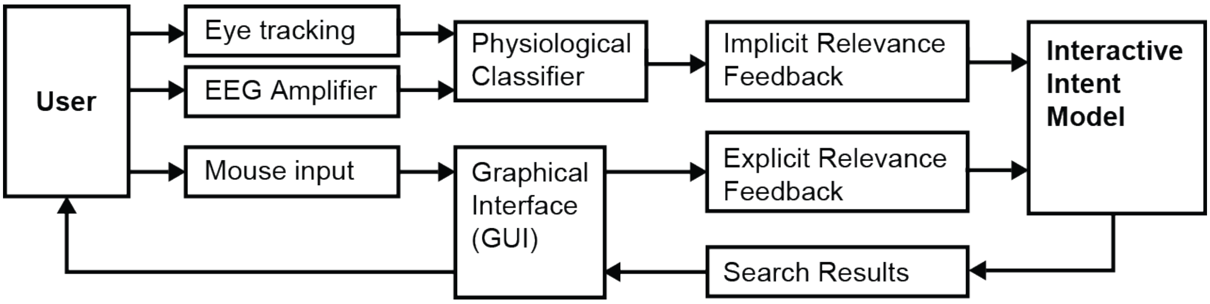

Figure 4. Summary of the system as a control loop during the online phase.

An Experiment in Neuroadaptive Literature Search

This experiment help 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 *Physiological 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 *Physiological 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 in order to associate the predicted relevance value to it. For collecting eye

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

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 dataset 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).

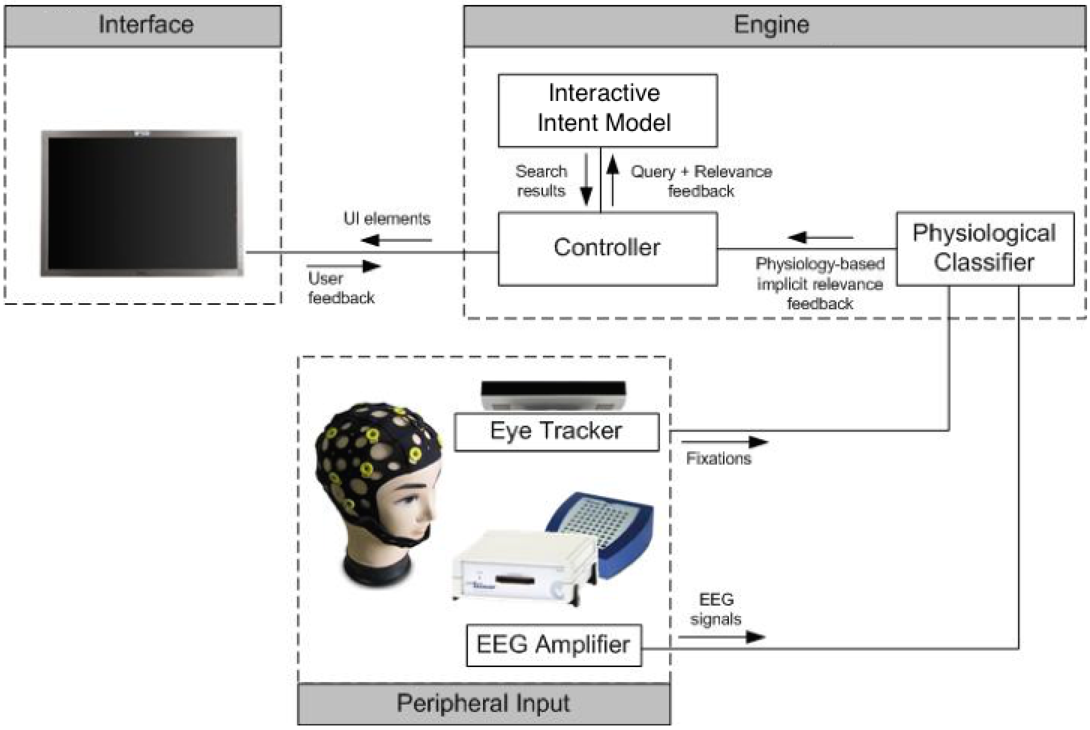

Figure 5. Components of the system.

Participants

Sixteen participants (3 females) took part in the experiment. The participants ranged from 22 to 39 years old (M = 28.3). Three participants were postdoctoral researchers,

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

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. While the physiological 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 physiological 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 physiological classifier, we generated a dataset that matched the application domain by using a subset of the dataset used by the interactive intent model system. The dataset 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 dataset. Upon selecting a topic, a series of keywords were shown to the user,

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

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 physiological classifier. [5]

The Online Phase. Once enough data had been collected and the physiological 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 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. Upon

[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.

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. In order 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 dataset 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, 2000). 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 the data presented to the user from which implicit relevance feedback was classified were generated in real-time.

The participants whose classification performance in the calibration phase was not significantly better than random were discarded from further analyses. Furthermore, participant *P05* 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.

For the remaining participants (N = 12), we studied how well the classification performance transferred to the online phase. In order to do so, 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.

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 physiological 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, in order 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, 397 documents were bookmarked, from which 277 were unique on the population level. We selected a subset of "representative" documents on the basis of bookmarked frequency. Documents were selected as "representative" for one of the system types (i.e., conditions) if on the population level, the document was bookmarked at least two more times than when using the other system type. This lead to a subset of 21 documents, which were 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, representing around 80% of the participants. On the population level, AUROC resulted in 0.61 ± 0.02 (mean ± standard error of the mean). Figure 6 presents the individual classification performances in the calibration phase.

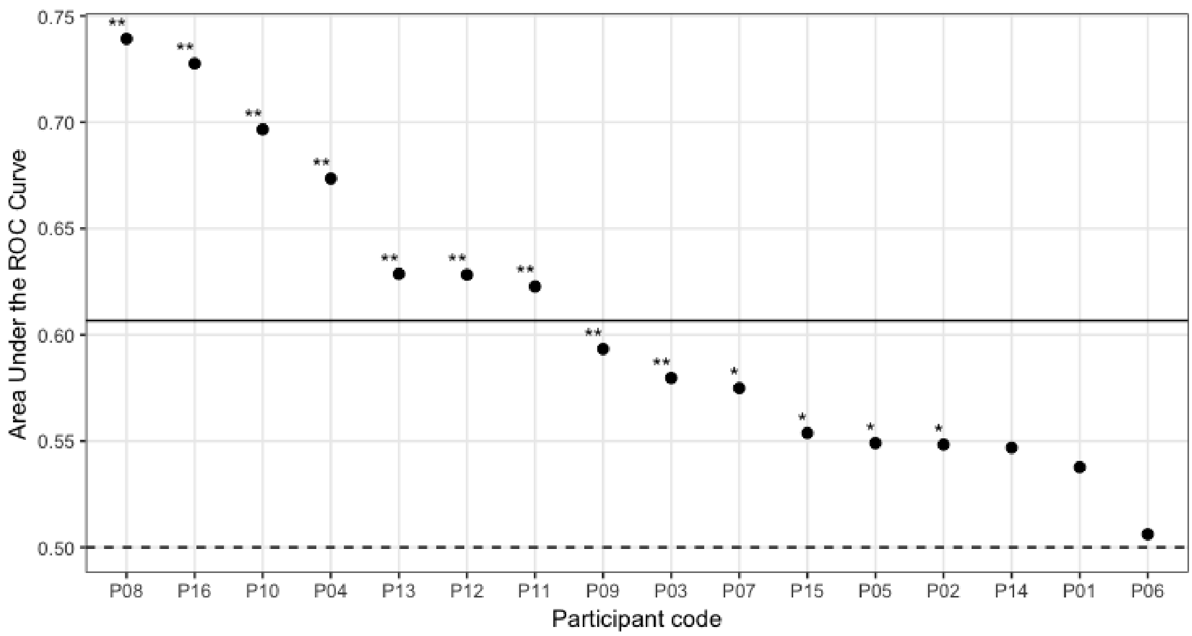

Figure 6. Individual classification performances in the calibration phase in terms of area under the ROC curve (AUROC), and improvement over the random baseline at the levels of $p < 0.05$ (*), and $p < 0.001$ (**). The horizontal lines represent the mean (solid) and random (dashed).

Online phase. Online relevance predictions as directly obtained through the physiological classifier presented averaged AUROC values on the population level of

0.53 ± 0.03 (mean $\pm$ 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 12 participants, representing over 80% of the 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.

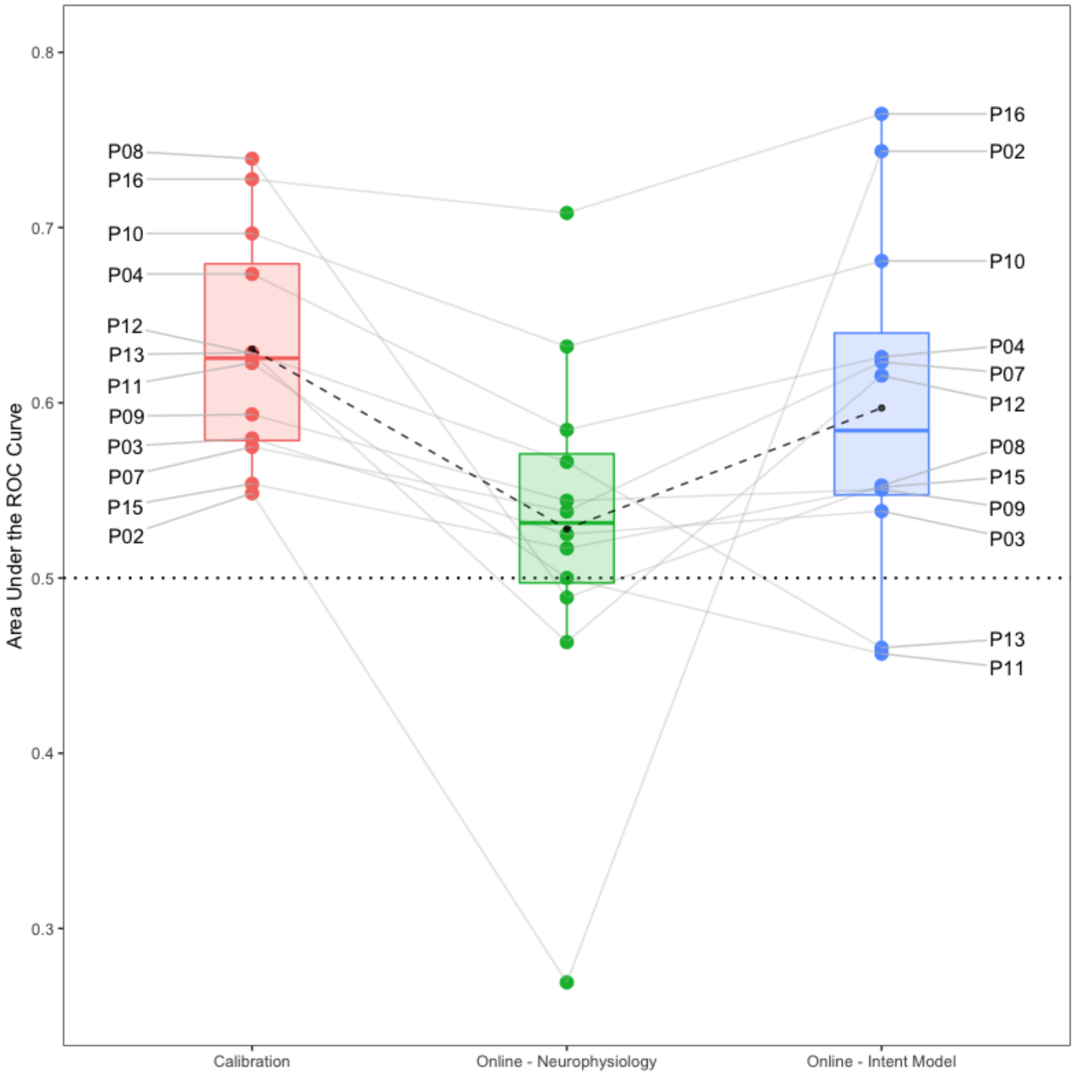

Figure 7. Individual classification performance in terms of area under the ROC curve (AUROC). Left: offline prediction in the "calibration phase". Middle: neurophysiological prediction in the "online phase". Right: intent model prediction in the "online phase". Smaller black dots and dashed lines indicate mean classification performance. The dashed horizontal line represents random classification.

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

This study indicates that we are able to reliably train classification models for implicit relevance prediction by using complex data domains and a computer science-related database. The results show that the classification performance significantly outperformed random predictions for over 80% 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 dataset of keywords was used.

In addition, the results show that the classification performances achieved using the controlled “calibration dataset” 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. While the classification performance decreased as expected, the overall distribution across participants remained above random classification levels.

Furthermore, 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.

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.

Our approach and study includes at least two limitations. The predicted relevance from physiology, while promising, still leaves room for improvement, both in terms of classification performance and uniformity across participants. Moreover the analysis on the selection behavior of bookmarked documents did not yield conclusive results in terms of task performance improvements yet. Future work should extend the presented results by further studying how the reported classification performances could transfer over to search task performance.

In conclusion the current work contributes showing that we can predict the relevance of keywords from neurophysiology with promising accuracy in a realistic search task and that this information can be integrated in a unified model in a IR system utilizing interactive intent modeling. Recently Mostafa and Gwizdka (2016) called for standardized practices in integrating BCI-based implicit feedback to IR for example discussing the need for standardizing search tasks in experiments. The proposed approach additionally contributes to discuss how to standardize the prediction of neurophysiology based relevance feedback.

Acknowledgements

We thank Mats Sjöberg, Antti Kangasräääsiö, 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, feb). Large-Scale Assessment of a Fully Automatic Co-Adaptive Motor Imagery-Based Brain Computer Interface. *PLOS ONE*, *11*(2), e0148886. doi: 10.1371/journal.pone.0148886

Agrawal, S., & Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *Icml (3)* (pp. 127–135).

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, USA: ACM. doi: 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, USA: ACM. doi: 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, USA: 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, USA: ACM. doi: 10.1145/2678025.2701389

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. doi: 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. doi: 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. doi: 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® in Human—Computer Interaction*, *9*(3-4), 151-308. Retrieved from `http://dx.doi.org/10.1561/1100000065` doi: 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, USA: ACM. doi: 10.1145/2856767.2856803

Eugster, M. J. A., Ruotsalo, T., Spapé, M. M., Barral, O., Ravaja, N., Jacucci, G., & Kaski, S. (2016, December). Natural brain-information interfaces: Recommending information by relevance inferred from human brain signals. *Scientific Reports*, *6*, 38580. doi: http://dx.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, USA: ACM. doi: 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, mar). Regularized Discriminant Analysis. *Journal of the American Statistical Association*, *84*(405), 165–175. doi: 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). Springer. doi: 10.1007/978-3-319-24917-9_8

Golenia, J.-E., Wenzel, M. A., Bogojeski, M., & Blankertz, B. (2017). Implicit relevance feedback from electroencephalography and eye tracking in image search. *Journal of Neural Engineering*. Retrieved from `https://doi.org/10.1088/1741-2552/aa9999` (accepted; open access)

Good, P. I. (2000). *Permutation tests : a practical guide to resampling methods for testing hypotheses* (2nd ed.). Springer.

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, USA: ACM. doi: 10.1145/2637002.2637011

Gwizdka, J., & Mostafa, J. (2015, jan). NeuroIR 2015: SIGIR 2015 Workshop on Neuro-Physiological Methods in IR Research. In *Acm sigir forum* (Vol. 49, pp. 83–88). ACM. doi: 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, New York, USA: ACM Press. doi: 10.1145/3020165.3022165

Jacucci, G., Fairclough, S., & Solovey, E. T. (2015, oct). Physiological Computing. *Computer*, *48*(10), 12–16. Retrieved from `http://ieeexplore.ieee.org/document/7310960/` doi: 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, USA: ACM. doi: 10.1145/2930238.2930243

Kauppi, J.-P., Kandemir, M., Saarinen, V.-M., Hirvenkari, L., Parkkonen, L., Klami, A., ... Kaski, S. (2015, may). Towards brain-activity-controlled information retrieval: Decoding image relevance from MEG signals. *NeuroImage*, *112*, 288–98. doi: 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, USA: ACM. doi: 10.1145/1148170.1148249

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

Ledoit, O., & Wolf, M. (2004, feb). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365–411. doi: 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, USA: ACM. doi:

10.1145/2484028.2484074

Mostafa, J., & Gwizdka, J. (2016). Deepening the role of the user: Neuro-physiological 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).

Nijholt, A., Tan, D., Pfurtscheller, G., Brunner, C., d. R. Millán, J., Allison, B., ... Müller, K. R. (2008, May). Brain-computer interfacing for intelligent systems. *IEEE Intelligent Systems*, *23*(3), 72-79. doi: 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, USA: ACM. doi: 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).

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., ... others (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).

Schäfer, J., & Strimmer, K. (2005, jan). A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*(1). doi: 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, oct). Real-time inference of word relevance from electroencephalogram and eye gaze. *Journal of Neural Engineering*, *14*(5), 056007. doi: 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` doi: 10.3389/fnins.2016.00023

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