A Comparison of Unimodal and Multimodal Models for Implicit Detection of Relevance in Interactive IR

Roberto González-Ibáñez
Departamento de Ingeniería Informática, Universidad de Santiago de Chile, Avenida Ecuador #3659 Estación Central, Santiago, Chile. E-mail: roberto.gonzalez.i@usach.cl

Aileen Esparza-Villamán
Departamento de Ingeniería Informática, Universidad de Santiago de Chile, Avenida Ecuador #3659 Estación Central, Santiago, Chile. E-mail: aileen.esparza@usach.cl

Juan Carlos Vargas-Godoy
Departamento de Ingeniería Informática, Universidad de Santiago de Chile, Avenida Ecuador #3659 Estación Central, Santiago, Chile. E-mail: juan.vargasgo@usach.cl

Chirag Shah
Department of Library and Information Science, School of Communication and Information, Rutgers University 4 Huntington Street, New Brunswick, NJ, 08901 USA. E-mail: chirags@rutgers.edu

Implicit detection of relevance has been approached by many during the last decade. From the use of individual measures to the use of multiple features from different sources (multimodality), studies have shown the feasibility to automatically detect whether a document is relevant. Despite promising results, it is not clear yet to what extent multimodality constitutes an effective approach compared to unimodality. In this article, we hypothesize that it is possible to build unimodal models capable of outperforming multimodal models in the detection of perceived relevance. To test this hypothesis, we conducted three experiments to compare unimodal and multimodal classification models built using a combination of 24 features. Our classification experiments showed that a univariate unimodal model based on the left-click feature supports our hypothesis. On the other hand, our prediction experiment suggests that multimodality slightly improves early classification compared to the best unimodal models. Based on our results, we argue that the feasibility for practical applications of state-of-the-art multimodal approaches may be strongly constrained by technology, cultural, ethical, and legal aspects, in which case unimodality may offer a better alternative today for supporting relevance detection in interactive information retrieval systems.

Received January 12, 2018; revised November 27, 2018; accepted December 26, 2018

 DOI: 10.1002/asi.24202

Introduction

Every day millions rely on search engines to find information online. Although decades of research and development have brought substantial transformations to these systems, users do not always see their information needs completely satisfied (Foresee, 2015). One possible cause for this problem can be attributed to the intrinsic limitations of search engines to understand the purpose of the searchers, their motivations, and their reactions to the information provided. A well-known problem in this regard is determining what information is perceived by users as relevant and what is not.

Searching information through search engines is often considered an interactive process (Marchionini, 2008). On the one hand, search engines rely on indexing and ranking algorithms to retrieve and present information to users in response to their queries. Users, on the other hand, react by exploring search results, evaluating information, saving information, reformulating their queries, formulating new ones, and finalizing their search session. Modern search engines and system-based search assistants may take advantage of this interactive process by looking at particular users' behaviors and personalization (Google, 2012). With this type of information, such systems may be able to infer users' intents, task type, task completion, and more important, whether or not users feel satisfied with the information provided by the system. Regarding the latter, one

way to infer satisfaction levels is by looking at users' perceived relevance. For doing so, explicit relevance feedback would be an appropriate approach; however, only a small percentage of users are usually willing to provide such information (Warner & Myer, 2003). In this sense, alternative methods for implicit detection of information relevance are required.

During the last decade, different studies on information retrieval (IR) have shown the feasibility to detect implicitly if a document is considered relevant by users. To tackle this problem, some have used single features—such as the time that users spent on documents (dwell time; Fox, Karnawat, Mydland, Dumais, & White, 2005; Guo & Agichtein, 2012) or click-through behaviors (Jung, Herlocker, & Webster, 2007)—to build classification models. Others, on the other hand, have built classification models using multiple features derived from behavioral, neurophysiological, and attentional measures, to name a few (Wang, Hawk, & Tenopir, 2000; Moshfeghi & Jose, 2013; Moshfeghi et al., 2013).

By definition, univariate models are unimodal models; however, multivariate models can be either unimodal or multimodal. In this context, a modality refers to a channel or source from where data are obtained. For example, mouse actions, facial expressions, electrodermal activity, and eye movements are obtained from different channels, thus different modalities. In this sense, a multivariate model that uses data from mouse movements and mouse clicks can be considered unimodal, whereas a multivariate model that combines data from mouse actions and eye movements can be regarded as multimodal.

It is typically expected that multivariate models (unimodal or multimodal), based on the combination of features derived from the interaction between users and systems, would lead to more accurate models of user behaviors. Although studies on implicit detection of information relevance have shown promising results when combining multiple features, it is not clear yet to what extent, if any, the use of multiple features, particularly those that lead to multimodal models, constitutes an effective and a practical approach compared to the use of univariate/unimodal models.

To address this research problem, we present a comparative study to contrast both unimodal and multimodal classification models of information relevance (as perceived by users). In particular, we hypothesize (H_1) that it is possible to build unimodal models capable of matching and even outperforming multimodal models in the detection of information relevance. To test this hypothesis we used data from a user study involving 45 participants facing a precision-oriented search task. In particular, we extracted 24 features from the interaction of users and webpages as part of their search process. These features include behavioral, expressive, attentional, and physiological measures. With these features, we built a set of classification models (both unimodal and multimodal) using logistic regression (LR), naive Bayes (NB), and support vector machine (SVM) to compare their performance.

Our methodological approach to building multimodal models is gradual; in other words, we generated a variety of models based on the combinations of features adding them one by one. To compare the classification models we used receiver operating characteristic (ROC) curves to compare the best unimodal and multimodal models built. Results from our classification experiments revealed that a univariate unimodal model based on the left-click action supports our hypothesis. In particular, we found that this model outperforms the best multimodal models built for this evaluation. On the other hand, our prediction experiment suggests that multimodality slightly improves early classification compared to the best unimodal models. Our findings have practical implications regarding the feasibility of implementing state-of-the-art multimodal models, which may be strongly constrained by technology, cultural, ethical, and legal aspects, in which case unimodality may offer a better alternative for supporting interactive information retrieval (IIR) systems.

The rest of the article is organized as follows. In the Related Work section, the background and related work are presented. Then the methodological approach used in this study is described in the Method section. The results are reported in the Results and Discussion section, with details and discussion on hypothesis testing. The article concludes with the Conclusion section, with an outline of the reported work, shortcomings, and pointers for future work.

Related Work

Traditionally, search engines determine the relevance of a document based on the correspondence between users' queries and the text contained in webpages (Brin & Page, 1998). Despite this interaction between users and system, users are often considered passive actors who issue queries to an IR system.

Several scholars have recognized that IR systems, and particularly search engines, can benefit from considering the interactive process between users and systems (for example, Marchionini, 2008). Users not only formulate and issue queries as passive actors in the search process, but they also have an active role when interacting with search results and information objects. This interaction process can be expressed through behavioral, cognitive, and affective responses. In this sense, it has been argued that search engines could increase satisfaction levels of users if such systems could identify the purpose of their searches, their motivations, and their reactions to the information provided (Teevan, Dumais, & Horvitz, 2010).

Although search engines typically focus on algorithmic and topical relevance, it has been recognized that information relevance can also be expressed through other dimensions such as cognitive, affective, and situational relevance (Saracevic, 1996). In addition, perceived relevance is a subjective component of information relevance that tells whether a given information object is considered relevant or not from the perspective of searchers. In this sense,

DOI: 10.1002/asi

perceived relevance may include situational, affective, and cognitive relevance.

Researchers have adopted different approaches to studying and using data derived from the interaction between users and search engines. Two common approaches are based on the use of a single (unimodality) and multiple (multimodality) sources of features (for example, behavioral, cognitive, and affective) to detect implicitly whether or not a document is relevant. These approaches have also been used to find out if searchers are satisfied or dissatisfied with the results provided by an IR system. Next, we present an overview of different studies that have used both approaches.

Some studies have shown that geographical location and time can influence how information relevance is determined (Baeza-Yates & Galleguillos, 2005). In fact, modern search engines consider location and related aspects, such as language and freshness, to retrieve and rank information that is relevant to the context of users.

Besides search context, researchers have paid attention to behavioral aspects of searchers as a result of the interaction with IR systems. Fox et al. (2005) looked into implicit measures that can be used to improve web search as an alternative to explicit ratings. Examples of this include the time spent on webpages (dwell time), clickthrough behaviors, and exit action. These particular measures were found to be strong predictors of user satisfaction. Along the same lines, Jung et al. (2007) focused on click data as an implicit source to provide relevance feedback to an IR system. The authors found that behaviors expressed through click actions in the interaction with an IR system, particularly with search results pages (SERPs), can inform what webpages are visited more frequently. Also, click actions in the interaction with information objects could provide feedback to IR systems beyond the scope of SERPs.

Others have linked search behaviors, such as click data, to affective aspects. For instance, Lopatovska (2011) studied search behaviors and their relation to emotions as expressed through facial expressions. According to the author, click actions (for example, left-click, scrolling) and specific facial expressions, such as happiness and surprise, are correlated. Recently, González-Ibáñez and Shah (2016) showed that smiles could act as implicit indicators of information relevance. More precisely, the authors found that smiles surround explicit actions of searchers such as bookmarking or saving snippets from webpages, thus implying that such documents were relevant. Moreover, the authors found that combining smiles and electrodermal activity (EDA) can be used to determine success in the completion of a search task. However, limited results were found when using affective features (that is, overall frequencies of facial expressions and EDA peaks) in univariate and multivariate classification models of information relevance.

Along the same lines, Arapakis, Jose, and Gray (2008) showed that affective responses (expressed through facial expressions and questionnaires) could be a valuable source of feedback to IR systems, which can be used to infer task difficulty and complexity. In a later study, Arapakis, Konstas, Jose, and Kompatsiaris (2009) showed that not only facial expressions but also other physiological signals such as EDA and skin temperature could be used to predict topical relevance. Using models based on SVM and nearest neighbor clustering (KNN), the authors found that the set of features based on facial expressions and peripheral physiological signals (used independently) outperform the accuracy of a baseline model.

Using a multimodal approach, Moshfeghi and Jose (2013) showed that dwell time in combination with affective and physiological signals could improve the implicit detection of information relevance. Also using a multimodal approach, Guo and Agichtein (2012) studied how mouse interactions in combination with dwell time can be used to detect whether an information object is relevant or not. As a result of this study, the authors found that the frequency and speed of mouse movements could indicate reading behaviors, and in turn, information relevance.

Finally, other studies have also investigated neurophysiological signals found through brain activity and eye tracking (Hardoon, Shawe-Taylor, Ajanki, PuolamÃ, & Kaski, 2007). Regarding brain activity, Moshfeghi et al. (2013) used functional magnetic resonance imaging (fMRI) to investigate areas of the brain that are activated when assessing topical relevance of images. According to the authors, the activations of three regions of the brain were found to differ when processing relevant and nonrelevant information. On the other hand, regarding eye tracking, Hardoon et al. (2007) showed that eye movements (using features such as fixations and saccades) could be used to infer queries as a result of the interaction with relevant documents. Also using eye tracking, Gwizdka and Zhang (2015) found that measures such as saccades duration, saccades length, fixation duration, and pupil dilation differ when searchers visit and revisit relevant and nonrelevant Wikipedia pages. Moreover, the authors found that pupil dilation can be related to cognitive aspects such as mental effort and attention in the context of relevant documents. Recently, González-Ibáñez, Escobar-Macaya, and Manriquez (2016) studied two mental states (that is, attention and relaxation) and blink strength—as measured by a low-cost EEG sensor—during the exposure of users to relevant and nonrelevant pages in a self-motivated search task. Results from this study showed significant differences between relevant and nonrelevant pages with respect to attention levels and blink strength.

Method

It is clear from the review of the literature that although capturing various physical, cognitive, and affective signals from a searcher could be useful in assessing the relevance of a document, we still lack a clear understanding of how individual features from these signals perform compared to a set of all available features. To address this research problem, we conducted controlled experiments using existing data from a user study on information search. The experiment

DOI: 10.1002/asi

was designed to test hypothesis H_1, which states that it is possible to build unimodal models capable of matching and even outperforming multimodal models in the detection of relevant information, as perceived by users.

Our methodological approach involves the following two main procedures: (a) data preprocessing and (b) evaluation. The first focuses on particular stages of the knowledge discovery in databases (KDD) life cycle and synchronization procedures for combining different data sources in the data set. The second consists of the process to build classification models using machine-learning techniques and the method for hypothesis testing. In the following subsections, we describe the data used in this study along with the data preprocessing and evaluation procedures.

Data

To carry out this study, we used the records of 45 single users from an existing data set (González-Ibáñez & Shah, 2014; González-Ibáñez & Shah, 2016). Note that although the original data set is larger, containing records from 135 users, 90 were part of pairs (dyads) who engaged in active text-based chat interactions as they performed a collaborative search process. Logs from these users were rather noisy given the purpose and type of features (for example, keystrokes, mouse actions, eye movements) investigated in this study. Thus, we only focused on the subset of single users. This data set includes search logs, users' actions, eye-tracking data, EDA, facial expressions, and questionnaire responses that were captured or generated using different tools. In particular, TechSmith Morae[1] for keyboard and mouse events, Mirametrix Viewer[2] for eye-tracking data, Affectiva Q Live[3] for EDA, an adapted version of Coagmento (González-Ibáñez & Shah, 2011) for questionnaires and web browsing activity, and FaceDetect (Sebe et al., 2007) for facial expressions classification. All data captured or generated with these tools were timestamped with a shared local computer's clock.

Regarding the search task, the participants in the study had to solve a set of *A Google a Day*[4] questions (for example, "Which U.S. vice president was able to read Greek, Latin and the world's second most commonly taught foreign language?"). This search task can be classified as fact-finding; however, multiple steps or queries are required to find the answers to these questions. As stated in the study from where the data set was obtained, in the ideal scenario all questions could be answered using two queries and only textual information.

Sessions in the original study lasted ~60 minutes. For this study, we used data from the main task, which lasted 25 minutes. During this task, participants were instructed to complete as many questions as they could, with a maximum of 5 minutes per question. In addition to that, participants were required to collect snippets from webpages that helped them to find the answer to each question. Then such webpages were classified as relevant.

Regarding sample demographics, 26 (57.78%) participants were women and 19 (42.22%) were men. Participants' age ranged between 18 and 27 years old. Moreover, most participants reported that their search skills were high to very high, with an average of 4.04 (SD = 0.69) on a five-point scale.

All the participants in the study were able to provide answers to at least five questions (M = 9.29, SD = 5.78). In the process, they were able to visit a total of 1,064 content pages, of which 419 were perceived as relevant. The average per participant was 23.64 (SD = 8.47) content pages and 9.52 (SD = 4.89) relevant pages.

Preprocessing

The first stage focused on preparing the data for the evaluation procedure. This stage started with the combination and synchronization of different data sets followed by particular stages of the KDD life cycle. The outcome of this process is a set of vectors of features, where each vector represents the interaction of users with each content webpage (relevant and nonrelevant webpages). The following sections describe each preprocessing stage.

Data synchronization and enrichment. As mentioned earlier, the data set comprises data captured or generated by different software/hardware products. Although one of these products (Coagmento) was adapted to record search logs and some behavioral data contextualized at the page level, the rest were general-purpose products. Hence, data were not contextualized in the search process. In this sense, this stage focused on combining and synchronizing all data sources to produce a single data set. Due to sampling rate differences between software and devices used in the study, synchronization algorithms were implemented. These algorithms were designed to associate visited webpages to keyboard events, mouse events, facial expressions, EDA, and eye-tracking data, keeping alignment differences as low as possible among the multiple data sources (which in all cases was kept under 1 second). Note that the synchronization process allowed marking each data source with the time frame in which each webpage was visited (the time when a user entered to a webpage and the time when she/he left it), thus providing context.

It is also important to note that as part of the enrichment process, specific events linked to users' behaviors and also physiological measures were converted into features. Examples of such features are EDA peaks, keystrokes combinations (for example, Ctrl+c and Ctrl+v), mouse events (for example, clicks and scrolling), facial expressions (for example, happy and surprised), and eye-tracking measures (for example, fixations and saccades).

Data selection. The second stage focused on selecting data for the study, which includes the selection of records

[1] https://www.techsmith.com/morae.html
[2] http://www.mirametrix.com/
[3] http://www.affectiva.com/
[4] http://www.agoogleaday.com/

DOI: 10.1002/asi

TABLE 1. List of features, their abbreviations, and modalities.

#	Feature	Abbreviation	Modality
1	Happy	Ha	Affective (expressive)
2	Sad	Sa	
3	angry	An	
4	surprised	Su	
5	EDA peak height	EDA	Affective (physiological)
6	condition	C	Experimental
7	Ctrl+f (find)	CF	Behavioral (keyboard actions)
8	Ctrl+a (select all)	CA	
9	Ctrl+c (copy)	CC	
10	Ctrl+v (paste)	CV	
11	left-click	LC	Behavioral (mouse actions)
12	right click	RC	
13	wheel scroll down	WD	
14	wheel scroll up	WU	
15	headSize+ (head close to screen)	HS+	Behavioral (body posture)
16	headSize– (head far from screen)	HS–	
17	dwell time	DT	Behavioral (reading attention)
18	number of fixations	NFx	Visual attention
19	fixations duration	FD	
20	saccades distance	SDi	
21	saccades duration	SDu	
22	saccades speed	SS	
23	number of blinks	NBl	
24	blinks duration	BD	

belonging to the main task (that is, horizontal selection) and the selection of features (that is, vertical selection). As part of the horizontal selection, only records associated with content pages were selected. On the other hand, in the vertical selection, 24 features were selected. The list of features and their modalities are presented in Table 1. Note that in this list, condition is a nominal variable, whereas all the other features are on a ratio level of measurement (that is, frequencies and averages). Moreover, although most features in this table are self-explanatory, there are a few that need further explanation. First of all, dwell time represents the time spent in webpages. Second, condition represents the group to which participants were assigned (that is, positive, negative, control) in the original study as part of the stimuli stage. Third, *headSize+* and *headSize–* represent how close and far a given participant was from the computer screen, respectively. Such movements could be used as proxies of mental states such as attention, focus, tiredness, and boredom, to name a few. These measures were calculated with respect to a threshold (a middle point) that was computed for each participant based on head movements captured through a webcam during the entire session. Finally, EDA peak height corresponds to the maximum height of EDA peaks that showed up during the interaction with webpages.

Data cleaning and coding. This stage focused on cleaning and coding the data to remove records with incomplete data (for example, missing facial expressions or eye-tracking data). In some cases, this led to discarding webpages from the final data set.

Regarding coding, some features were recoded and normalized. For instance, a binary representation for expressing the presence (1) and absence (0) was used for affective and the majority of behavioral features. To avoid noisy data, a criterion similar to that described in Arapakis et al. (2008; that is, a threshold to filter facial expressions classified by a software) was applied. As a result, facial expressions were coded as present only if their probability of occurrence was 0.8 or higher. Otherwise, facial expressions were coded as absent. Dwell time and eye-tracking features were not recoded.

Finally, the data set was synthesized for the data mining stage by producing a set of vectors with the features and class (that is, relevant and nonrelevant) representing each webpage. For doing so, features were expressed or aggregated in terms of frequencies (facial expressions, EDA peaks, keystrokes, and mouse events), average values (eye-tracking measures), raw values (dwell time), and nominal values (condition).

Data summary. As a result of this preprocessing stage, the final data set contained records from 958 webpages, of which 318 (33.19%) were relevant and 640 (66.81%) were not relevant. Finally, to obtain a balanced data set, 318 nonrelevant pages were randomly selected; hence, records linked to 636 webpages were used in the following analyses.

Although our data set can be considered small when compared to large data sets typically used in IR studies, it is important to consider that the set of webpages used in this study came from a complex lab study. The study involved multiple instruments (that is, webcam, eyetracker, EDA sensor, mouse and keyboard logger, and screen capture software) that were used to collect data from each user. Conducting such a study with more users would not be feasible at this point (especially when it comes to using the same technology that is no longer available or supported) and even at the time the study was conducted. Despite the size of the data set, each page is linked to hundreds or thousands of records captured with different sources. For instance, a typical stay of 30 seconds on a webpage produced 1,800 eye-tracking data points (at 60 Hz), 900 EDA data points (at 30 Hz), and 450 video frames (at 15 fps). Although dwell time varied from page to page, with an average of 15 seconds per page, an average of 572,400 eye-tracking data points, 286,200 EDA data points, and 143,100 video frames were captured and analyzed. In addition, other sources of data such as keystrokes and mouse actions were also collected, analyzed, and used to train our classification and prediction models.

Evaluation Procedure

The second stage focused on the process to build classification and prediction models using machine-learning

DOI: 10.1002/asi

techniques and hypothesis testing. Models were built in the following three experiments that focus on different sets of features and purposes: (a) classification models using experimental, behavioral, and affective features; (b) classification models using experimental, behavioral, affective, and eye-tracking features (that is, all features); and (c) prediction models.

In the first experiment, the focus was on building classification models that use features that can be captured with technology commonly available to users (17 features). The second experiment uses these 17 features plus a set of seven features captured with technology that is rarely available to the lay user (that is, eye tracking). Finally, in the third experiment, the main purpose was to build classification models at different timepoints of participants' exposure to both relevant and nonrelevant documents. This experiment aimed to perform timely classification (prediction) whether a page will be perceived as relevant or nonrelevant by searchers.

Following this, we built unimodal classification models (both univariate and multivariate) using NB, LR, and SVM over the group of selected features. In the first experiment, an incremental approach for combining features (from 2–7) was followed to build multimodal models.

Hypothesis testing, on the other hand, focused on testing H_1. To compare the classification models we used ROC curves and statistical tests to compare the best unimodal model, the best multimodal models, and a baseline model.

Results and Discussion

Below we present our results according to the evaluation procedure described in the previous section. First, we focus on the experiments with different classification models. Then we present results corresponding to hypothesis testing.

Classification Experiments

Next, we present results of the three experiments conducted to build and compare unimodal and multimodal models in classification and prediction tasks.

Experiment 1: Experimental, behavioral, and affective features. In this first experiment, we focused on experimental (that is, condition), behavioral (that is, dwell time, head movements, mouse actions, and keystrokes), and affective features (that is, facial expressions and EDA peaks). Overall, 17 features were considered to build classification models.

Feature selection. First, we conducted a feature selection procedure to evaluate the contribution of the different features in the distinction of relevant and nonrelevant webpages. To perform this procedure, we used built-in algorithms as implemented in Weka 3,[5] in particular, χ^2 as evaluator and Ranker as a search method.

[5] http://www.cs.waikato.ac.nz/ml/weka/

TABLE 2. Results for selected features (7 of 17) using χ^2 and Ranker.

Rank	Feature	Ranked-Weka
1	left-click (LC)	374.511
2	headSize+ (HS+)	117.124
3	dwell time (DT)	116.218
4	headSize– (HS–)	65.695
5	Ctrl+c (CC)	10.359
6	happy (Ha)	8.102
7	surprised (Su)	8.102

Using all 17 features for building all possible classification models for our gradual approach would derive a total of 131,071 models. However, as shown in Table 2, only seven features were ranked high. Note that the behavioral feature left-click was ranked in the highest position, which may reflect the active interaction of users with relevant webpages in contrast to nonrelevant ones in the context of the search task. Indeed, left-clicks were found in all relevant pages. Conversely, left-clicks were found in 265 (83.33%) irrelevant pages. On average, users interacted with relevant pages through left-clicks 5.22 times, whereas this number drops to 1.81 times for irrelevant pages. The remaining 10 features were ranked with zero value.

Models. Overall, we built 138 classification models with each machine-learning approach (that is, NB, LR, and SVM). First, we built models using the seven selected features, in particular, seven univariate unimodal models, two multivariate unimodal models (that is, [HS+, HS–] and [Ha, Su]), and 118 multimodal models using our gradual approach for combining these features. Second, we included 10 univariate unimodal models with the remaining features and one multimodal model that combined all 17 features. Each model was trained and tested using 10-fold cross-validation.

For the three classification approaches, dwell time was used as a baseline model (Fox et al., 2005; Guo & Agichtein, 2012). Moreover, due to the large number of models built, we only present results for the top-5 models and the baseline model. Note that all classification models (including those not reported here) were ranked in terms of accuracy.

As shown in Table 3, the highest accuracy (63.679%) for the baseline models (using dwell time as a unique feature) was reached by LR, followed closely by SVM and NB. Note that precision, recall, and F-measure for these models were very similar. Although dwell time has been reported and used as a predictor of user satisfaction (Fox et al., 2005) and page utility (White & Huang, 2010), the results for accuracy using this feature were, on average, 18.124% below the best models built using LR, NB, and SVM. Note that dwell time in both relevant and nonrelevant pages (including outliers) varied within similar ranges, as illustrated in Figure 1. It is also important to mention that dwell time might be affected as a result of the time constraints used in the original study from where these data was obtained (that is, a maximum of 5 minutes per question and 25 minutes for the entire search task).

DOI: 10.1002/asi

TABLE 3. Results for Experiment 1: Baseline and top-5 models.

Classifier	Features	# of features	# of modalities	Accuracy	Precision	Recall	F-Measure
LR	Baseline (DT)	1	1	63.679%	0.641	0.637	0.634
	1 to 17	17	7	**83.333%**	0.834	0.833	0.833
	CC,LC,HS+	3	3	81.447%	0.817	0.814	0.814
	Ha,Su,DT,CC,LC	5	4	81.446%	0.817	0.814	0.814
	Ha,DT,LC,HS–	4	4	81.446%	0.816	0.814	0.814
	Ha,DT,CC,LC,HS–	5	5	81.446%	0.816	0.814	0.814
NB	Baseline (DT)	1	1	63.364%	0.671	0.634	0.612
	HS+,LC	2	2	**74.213%**	0.756	0.742	0.739
	CC,LC,HS–	3	3	74.213%	0.760	0.742	0.738
	LC,HS–	2	2	73.899%	0.762	0.739	0.733
	CC,LC,HS+	3	3	73.742%	0.752	0.737	0.734
	LC	1	1	73.427%	0.765	0.734	0.726
SVM	Baseline (DT)	1	1	63.522%	0.640	0.635	0.632
	LC	1	1	**87.421%**	0.885	0.874	0.873
	Su,LC	2	2	87.421%	0.885	0.874	0.873
	LC,HS–	2	2	87.421%	0.885	0.874	0.873
	CC,LC,HS–	3	3	87.421%	0.883	0.874	0.873
	CC,LC,HS+	3	3	87.421%	0.883	0.874	0.873

Regarding multimodal models, it was found that using all 17 features led to the highest accuracy (83.333%) for LR. Nevertheless, for NB and SVM the 17-features models did not show up in the top-15 models. On the other hand, the results for SVM revealed that the best models included one univariate unimodal model based on the left-click feature (87.421%). More important, the top-21 multimodal models (using LR, NB, and SVM) included left-click as one of the contributing features; however, for the particular case of SVM, results for these multimodal models did not outperform the left-click classification model. This finding is consistent with results reported in our feature selection stage, in which the left-click presented the highest rank score. Moreover, this result suggests that the best unimodal model could not be improved by adding more features, whether to build multivariate unimodal models or multimodal models.

To evaluate the contribution of the left-click feature in the multimodal models, we investigated all multimodal models that did not include this feature. In this evaluation, we added the univariate unimodal classification models based on the left-click feature as an additional baseline (Baseline 2) for each classification approach. As shown in Table 4, such baseline models reported the highest accuracy compared to Baseline 1 (based on left dwell time) and all multimodal models that did not include the left-click feature. Note that in this table we report results for the models with the highest accuracies after Baseline 2. We

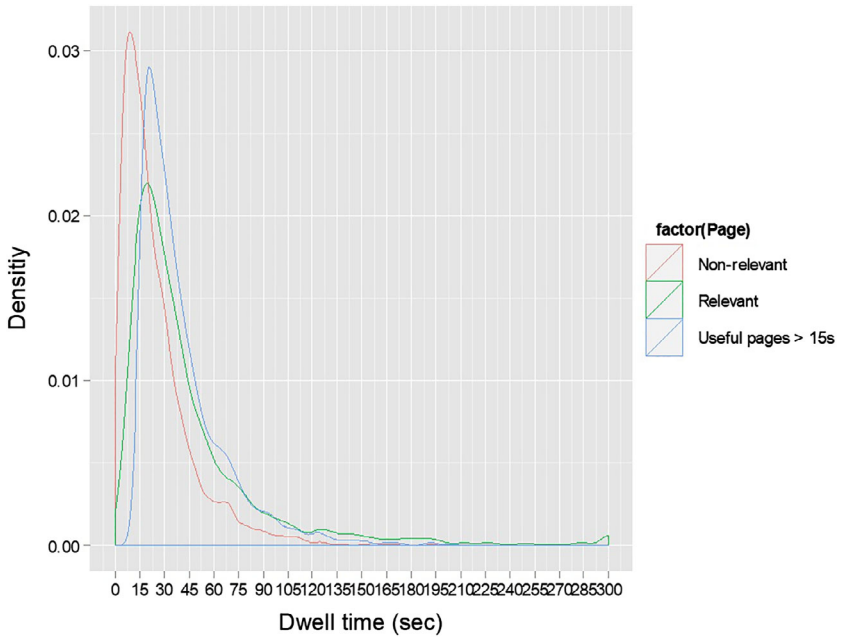

FIG. 1. Density plot for webpages (nonrelevant, relevant, useful) dwell time. [Color figure can be viewed at wileyonlinelibrary.com]

DOI: 10.1002/asi

TABLE 4. Results for Experiment 1: Multimodal models without using the left-click feature.

Classifier	Features	# of features	# of modalities	Accuracy	Precision	Recall	F-Measure
LR	Baseline 1 (DT)	1	1	63.679%	0.641	0.637	0.634
	Baseline 2 (LC)	1	**1**	**78.930%**	0.795	0.789	0.788
	DT,HS+	2	2	67.138%	0.680	0.671	0.667
	CC,HS+	2	2	65.408%	0.663	0.653	0.649
	Ha,Su,DT,CC	4	3	64.937%	0.654	0.649	0.647
	Ha,DT,CC,HS–	4	4	64.465%	0.649	0.645	0.642
	Ha,DT,HS–	3	3	63.836%	0.642	0.638	0.636
NB	Baseline 1 (DT)	1	1	63.364%	0.671	0.634	0.612
	Baseline 2 (LC)	1	**1**	**73.427%**	0.765	0.734	0.726
	CC,HS+,HS–	3	2	64.465%	0.675	0.645	0.629
	CC,HS+	2	2	62.421%	0.666	0.624	0.599
	HS+	1	1	61.949%	0.664	0.619	0.592
	CC,HS–	2	2	59.119%	0.617	0.591	0.567
	HS–	1	1	58.490%	0.615	0.585	0.556
SVM	Baseline 1 (DT)	1	1	63.522%	0.640	0.635	0.632
	Baseline 2 (LC)	1	**1**	**87.421%**	0.885	0.874	0.873
	CC,HS+,HS–	3	2	67.138%	0.681	0.671	0.667
	CC,HS+	2	2	64.779%	0.659	0.648	0.642
	HS–	1	1	62.264%	0.638	0.623	0.612
	CC,HS–	2	2	61.635%	0.630	0.616	0.606
	Su	1	1	52.515%	0.594	0.525	0.419

also present the accuracy of models listed in Table 3 without the contribution of the left-click feature. On average, the accuracy of the LR, NB, and SVM multimodal models including the left-click feature reported in Table 3 were 16.784% (*SD* = 0.671), 13.521% (*SD* = 2.037), and 27.122% (*SD* = 5.363) higher than the corresponding models without this feature (Table 4), respectively. This indicates a clear contribution of the left-click feature in the high performance of the unimodal and multimodal models reported in Table 3.

Excluding the three univariate unimodal models based on the left-click feature, our results showed that unimodal models (both univariate and multivariate) were among the ones with the lowest accuracy. In particular, the worst model was the one using the wheel scroll-up feature with LR (46.855%).

These results illustrate that even by combining features from different modalities, classification models using different machine-learning approaches do not necessarily improve in terms of accuracy and other measures such as precision and recall. Moreover, it was shown that left-click, an event that can be easily captured in desktop computers and laptops, can be a good indicator of information relevance in IIR so that it could be used as an implicit measure in relevance feedback.

Experiment 2: Experimental, behavioral, affective, and eye-tracking features. In our second experiment we used the 17 features included in the first experiment plus seven features obtained from eye-tracking data (Table 1). This experiment was conducted to study the contribution of features extracted from state-of-the-art technology in the implicit detection of information relevance. Note that due to noise and loss in eye-tracking data—as a result of sensor accuracy, calibration, and/or participants' freedom to move their heads during the study—we conducted an experiment independent from the first one. Unlike the approach to combine features incrementally in the first experiment, here we focused on comparing four models, namely, (a) our baseline model based on dwell time, (b) a model that combines the seven eye-tracking features, (c) a model that combines the most prominent features among all features, and (d) a model that combines all 24 features.

Feature selection. As in the first study, we performed an additional feature selection procedure with all 24 features. The goal of this analysis was to determine the contribution of the different features in the distinction of relevant and nonrelevant webpages. We used built-in algorithms of Weka 3, in particular, χ^2 as evaluator and Ranker as a search method.

As shown in Table 5, only nine features were ranked high. Once again, the behavioral feature left-click was ranked at the highest position. From the list of seven eye-tracking features, only two of them appeared among the nine features. Because the remaining 15 features were ranked with zero value, they were discarded. Unlike the first experiment, in this one we only sought to evaluate the performance of unimodal and multimodal models

TABLE 5. Results for selected features (9 of 24) using χ^2 and Ranker.

Rank	Feature	Ranked-Weka
1	left-click (LC)	374.511
2	headSize+ (HS+)	117.124
3	dwell time (DT)	116.218
4	headSize– (HS–)	65.695
5	num. of fixations (NFx)	55.663
6	num. of blinks (NBk)	18.938
7	Ctrl+c (CC)	10.359
8	happy (Ha)	8.102
9	surprised (Su)	8.102

DOI: 10.1002/asi

TABLE 6. Results for Experiment 2 using eye-tracking features.

Classifier	Features	# of features	# of modalities	Accuracy	Precision	Recall	F-Measure
LR	Baseline 1 (DT)	1	1	63.679%	0.641	0.637	0.634
	Eye-tracking features	7	1	62.106%	0.629	0.621	0.615
	Selected features	9	6	64.465%	0.649	0.645	0.642
	All	24	8	**83.333%**	0.833	0.833	0.833
NB	Baseline 1 (DT)	1	1	63.364%	0.671	0.634	0.612
	Eye-tracking features	7	1	60.220%	0.634	0.602	0.577
	Selected features	9	6	**63.993%**	0.677	0.640	0.620
	All	24	8	63.522%	0.662	0.635	0.619
SVM	Baseline 1 (DT)	1	1	63.522%	0.640	0.635	0.632
	Eye-tracking features	7	1	61.635%	0.645	0.616	0.596
	Selected features	9	6	**86.006%**	0.865	0.860	0.642
	All	24	8	85.849%	0.859	0.858	0.858

involving eye-tracking features. Hence, we trained and tested models with specific combinations of features.

Models. As a result of this study, we built three classification models with each machine-learning approach (that is, NB, LR, and SVM). In particular, we built multivariate unimodal models using a combination of the seven eye-tracking features, multimodal models using the nine selected features, and multimodal models using all 24 features. As in the first experiment, we included baseline unimodal models using only dwell time as a feature. Moreover, each model was trained and tested using 10-fold cross-validation.

As shown in Table 6, none of the multivariate unimodal models (based on the seven eye-tracking features) outperformed the baseline models based on dwell time. When combining all features, the results for accuracy were on average 14.046% (*SD* = 12.101) higher than the baseline models. However, the accuracy of the NB model using all 24 features (that is, eight modalities) was very similar to the respective baseline model.

The best classifier in this experiment was achieved through SVM using the nine selected features. Note that this model included the left-click feature; however, its accuracy (86.006%) did not outperform that of the best univariate unimodal model (that is, SVM + left-click) and the best 21 multimodal models (using fewer than seven features), which was 87.421%.

Moreover, when looking at the three models of the previous experiment that share the seven selected features (that is, left-click, headSize+, dwell time, headSize–, Ctrl +c, happy, and surprised), it was found that the two selected eye-tracking features worsen the classifiers' accuracies. In particular, the accuracy of the seven-feature model using LR was 80.346% and it was ranked in the 47th position. On the other hand, the accuracy of the seven-feature model using NB was 64.308% and it was ranked in the 55th position. Finally, the accuracy of the seven-feature model using SVM was 86.793% and it was ranked in the 49th position.

Experiment 3: Prediction of relevance judgments. The third experiment focused on building and comparing classification models to perform timely predictions of users' relevance judgments. In this context, prediction was tackled as a classification task at different timepoints during the exposure to information objects (that is, webpages). Specifically, we contrasted the performance of the best unimodal model, the best multimodal model, and a baseline model in the binary classification of information relevance at the following timepoints of exposure to information objects: 25%, 50%, 75%, and the moment right before leaving web documents (T_f-1).

To prepare data for this experiment, we first normalized the time spent in pages to homogeneously estimate the different timepoints (that is, 25%, 50%, 75%, T_f -1). Second, we generated vectors for the different timepoints with the features selected according to the procedure described below. Note that each vector represents a partial stay on a page. Third, we added the classes (relevant and not-relevant) to each vector based on the explicit judgment provided by the participants at T_f (page exit time) to denote whether a page was relevant page to them. Finally, training and testing sets were defined automatically as part of our 10-fold cross-validation approach.

The selection of a unimodal model was based on the results reported in the previous experiments. Because our best unimodal model was built with SVM using the left-click feature, we then proceeded to use SVM and this feature to build classification models at the four timepoints stated above. Conversely, to select the best multimodal model, we had to analyze a set of 21 models built in the previous experiments that fit in this category. These models achieved the same accuracy, precision, recall, and F-measure, yet they operated with different sets of features. In this regard, it is noteworthy that in all these 21 models, the left-click feature was involved. Furthermore, some of these models included affective features based on facial expressions and also behavioral features based on head movements and keyboard actions.

To select one model among these 21 instances, we compared their performance at the four timepoints (that is, 25%, 50%, 75%, T_f-1). First, we evaluated and contrasted their performance at 25% of exposure time. Then, if more than one model classified as the best model at this point, we used the features comprised in these models to build and compare new models at 50%, and so on, until identifying only one model with the best performance. As a result

DOI: 10.1002/asi

TABLE 7. Results for Experiment 3: Baseline and best models at four exposure moments.

Models with SVM	Features	# of features	Exposure time			
			Accuracy at 25%	Accuracy at 50%	Accuracy at 75%	Accuracy at T_f-1
Baseline	Page elapsed time	1	47.38%	47.70%	48.01%	60.70%
Best unimodal	LC	1	57.69%	66.72%	76.23%	**87.32%**
Best multimodal	LC, CC, HS+, HS–	4	**67.04%**	**67.99%**	**78.13%**	87.16%

of this process, we found that the best multimodal model was built with SVM over four features, namely, left-click (LC), Ctrl+C (CC), headSize+, and headSize−.

Regarding the baseline model, we built one model for each timepoint. Specifically, we built SVM models using the actual time spent on the webpage at the four timepoints. We refer to this feature as *pageElapsedTime*, which is an incremental version of the dwell time feature that was used in the previous experiments.

Table 7 and Figure 2 present a summary of the results for these models (that is, baseline, best unimodal, and best multimodal) at the four points of exposure to information objects. According to these results, the partial time spent in a webpage (that is, *pageElapsedTime*) was a poor predictor of perceived relevance. Indeed, the highest accuracy of the baseline models was 60.70% right before leaving webpages (that is, at T_f-1), which is rather close to the actual dwell time. The accuracy of such models in previous timepoints did not exceed chance levels (50%).

Regarding the best multimodal model, an interesting finding from this experiment is the contribution of features from multiple sources to enhance the classification accuracy. As shown in Table 7, the accuracy of this model at 25% of exposure time was 67.04%, which is higher than the accuracy reached by the baseline model at T_f-1 and almost 10% higher than the accuracy reached by the best unimodal model. At 50% and 75% of exposure, the accuracy of the best multimodal model increases; however, the gap with the best unimodal model tended to decay. Finally, at T_f-1, the best unimodal model outperformed the best multimodal model.

By taking into account the study design (that is, task, time constraints, and so on), these findings suggest that multimodality could have a positive impact in predicting perceived relevance at early stages of exposure to information objects.

Hypothesis Testing

To test hypothesis H_1, we used ROC curves and statistical tests to compare the best unimodal model, the best multimodal model, the best multimodal model that does not use the left-click feature, and a baseline model. To perform this

TABLE 8. AUCs and accuracies for baseline based on dwell time (DT), best unimodal, best multimodal, and best multimodal without left-click (LC).

Model	AUC ROC	Accuracy (%)
Baseline (DT)	0.738	63.84
Best unimodal	**0.902**	87.42
Best multimodal	0.842	87.58
Best multimodal w/o LC	0.749	67.45
Reference line	0.500	—

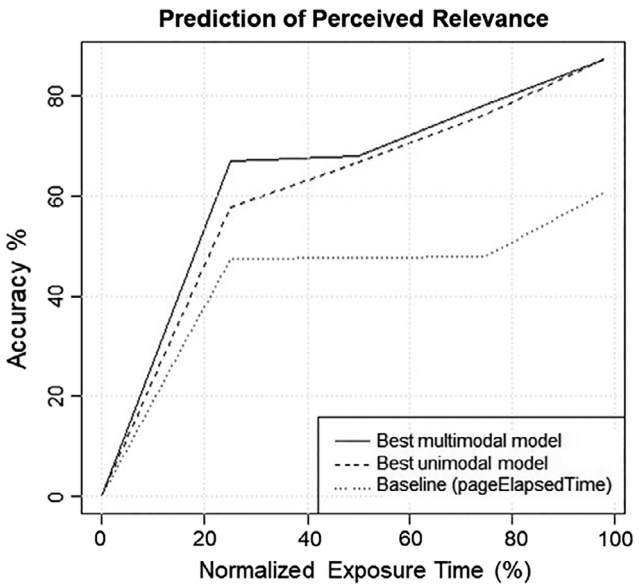

FIG. 2. Comparison of models' prediction accuracies at different exposure times.

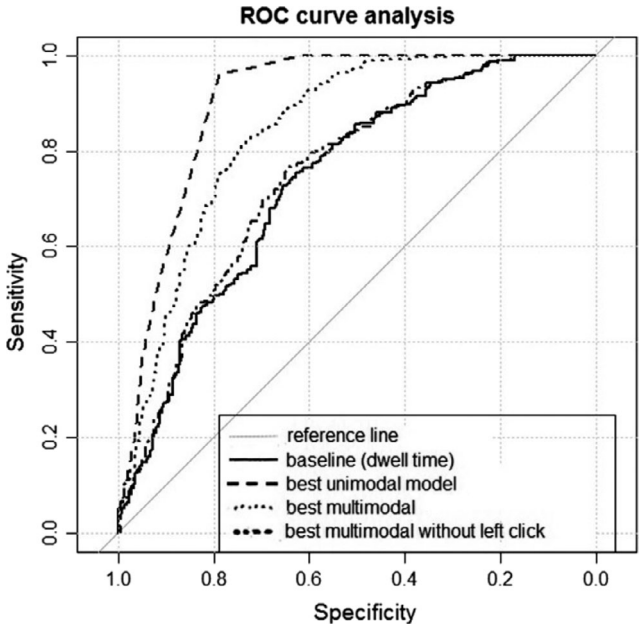

FIG. 3. ROCs for best unimodal model, best multimodal model, best multimodal model without left-click feature, and baseline model based on dwell time. All models built with SVM.

DOI: 10.1002/asi

TABLE 9. Results for hypothesis testing of the AUC of ROCs.

		Baseline model	Best unimodal	Best multimodal	Best multimodal w/o LC
Reference line	D	−2.5509	−4.3702	−3.6957	−2.6778
	df	33.9060	32.2310	32.8710	33.8090
	p-value	.01543*	.00012**	.00079**	.01135*
Best multimodal w/o LC	D	0.4255	−6.6241	−3.7230	—
	df	1269.6510	1105.4950	1222.4830	—
	p-value	.67050	5.443e11**	.00021**	—
Best multimodal	Z	−8.9361	4.8019	—	—
	p-value	2.2e-16**	.000002**	—	—
Best unimodal	Z	−8.3885	—	—	—
	p-value	2.2e-16**	—	—	—

Z = statistic for DeLong method; D = statistic for Bootstrap method.
$^{*}p < .05$; $^{**}p < .01$.

evaluation, we used the *pROC* and *RWeka* packages for *R-Project x64 3.1.2*. Because of the results obtained in the above experiments, the best classification models were chosen only from the first experiment using the area under the ROC curves (AUC) and accuracy as selection criteria. As a result, three SVM models were selected. First, a univariate unimodal model based on the left-click feature. Second, a multimodal model based on the left-click, Ctrl+c (copy action), headSize+, and headSize− features. Third, a multimodal model based on the Ctrl+c, headSize+, and headSize+. In addition, we included the SVM baseline model based on dwell time and a reference line with AUC equal to 0.5.

As shown in Table 8 and depicted in Figure 3, the model with the lowest AUC corresponds to the baseline model based on dwell time (AUC = 0.7381). Similarly, the best multimodal model that does not include the LC feature is slightly better than the baseline model (AUC = 0.7498). Yet both models are better than chance. On the other hand, the model with the highest AUC is a unimodal one based uniquely on the LC feature (AUC = 0.908). Note the that results of this evaluation (using R-Project x64 3.1.2) presents slight differences with respect to those obtained in Weka. Such differences could be attributed to implementation differences.

To test whether the AUC of ROCs are significantly different from each other, we used the *roc.test* function of the *pROC* package. As reported in Table 9, the AUC of ROC of the best unimodal model is significantly higher ($p < .01$) than those of the other models and the reference line. These results support our hypothesis H_1, in other words, it is possible to build unimodal models capable of matching and even outperforming multimodal models in the detection of information relevance, as perceived by users.

In addition, our results also indicate that our best univariate unimodal model could not be improved by adding additional features. In fact, the best multimodal model, which includes this particular feature, was significantly lower than our best univariate unimodal model.

Conclusion

In the investigation reported in this article, we compared two common approaches for implicit detection of information relevance (that is, unimodal and multimodal). Three experiments to build models and a formal procedure to compare the best models were conducted to accomplish this goal. For the particular case of multimodal approaches, previous studies have shown that the combination of features derived from the interactive process with IR systems could contribute to identifying when information is considered relevant by searchers. In spite of the potential benefits of multimodality, we focused on studying to what extent this is an effective approach compared to unimodality. In particular, we hypothesized that one could build unimodal models capable of matching and even outperforming multimodal models in the detection of information relevance, as perceived by users.

Our results from Experiments 1 and 2 showed that multimodal models could lead to high performance in the classification of relevant and nonrelevant documents. However, it was found that the performance of the best multimodal models was highly influenced by the presence of the left-click feature. More important, univariate unimodal models based on this feature reported high levels of accuracy for LR and NB, and the highest accuracy when using SVM. Our analyses also revealed that combining the left-click feature with features from the same and other modalities did not improve performance. Further analysis showed that our best univariate unimodal model (that is, SVM + left-click feature) was significantly better than the best multimodal models (with and without this feature); thus, our hypothesis was supported.

The positive influence of the left-click feature in the classification performance could be strongly related to the type of task (that is, fact finding) performed by the participants in the study from where data were obtained. It could be the case that when looking for particular facts, mouse actions tend to be more frequent than in other types of search task (for example, exploratory search). This behavioral phenomenon could be related to the level of interactions required to locate specific pieces of information (for example, highlight fragments of text, clicking on the scroll bar to skim the document).

Although the results for multimodality were not positive in Experiments 1 and 2, our third experiment revealed that multimodality could contribute when attempting to perform

DOI: 10.1002/asi

timely classification of information relevance. Indeed, our results suggest that when combining features from different sources at early stages of exposure to an information object, classification accuracy can improve. However, this effect tends to disappear when the user is close to abandoning the webpage.

It is important to consider that the extraction and usage of features during the search process may be affected by a variety of factors that could limit the feasibility to build IIR systems supported by such features. Examples of these factors include culture, privacy, ethics, laws, technology, and cost. Regarding the latter, regardless of the use of unimodal or multimodal approaches to implicitly detect whether a document is relevant or not, the cost (for example, computational, economic, and so on) may depend directly on the type of modalities involved and the particular technologies required to capture data from such modalities, among other factors. For instance, a unimodal model based on facial expressions (for example, happy, sad, and so on) could imply a higher computational and technical cost compared to a multimodal model involving behavioral features obtained from keystrokes and mouse actions.

In addition to computational power and technical challenges, it is important to consider ethical and legal limitations of implementing classification models with particular modalities. For example, although some studies have shown that expressive features such as facial expressions and head movements can contribute to the classification of information as relevant and nonrelevant, the identification of such features requires not only technological capabilities but also the explicit consent of users to allow systems to access capturing devices and data. Moreover, data captured by such devices (for example, video or images with users' faces) must be processed on the client or server side to extract the features of interest; hence, someone must assume that computational cost. Although some users may be willing to allow access to their capturing devices to improve their search experience, others may hesitate due to privacy, cultural, or ethical concerns.

Along the same lines, although web cameras are common devices for users, more sophisticated technology such as eye trackers and other neurophysiological sensors (for example, EEG, ECG) are uncommon outside of lab settings. In recent years, with the constant evolution of technology, low-cost sensors are becoming more popular among people. It may take some time before they are widely adopted; however, even after a wide adoption, their practical usage by IIR systems may be still constrained by the same factors described above. Moreover, as shown in this study, features provided by such technology may not necessarily contribute to a better classification of information as relevant or not.

The contribution of studies using multimodal approaches, especially those involving the use of neurophysiological sensors and sophisticated techniques to understand underlying mental aspects linked to information processing is unquestionable; however, it is also important to consider not only long-term but also mid- and short-term application possibilities. In this sense, although multimodality may outperform unimodality in the implicit detection of information relevance, we have also shown that multimodality is not necessarily an effective alternative to unimodality. In fact, we showed that classification models using a single mouse action (left-click) could be as effective as the best multimodal model built. More important, we also showed that multivariate models (unimodal or multimodal) that include the left-click feature does not necessarily lead to better classification models; conversely, in some cases it worsen the results.

Although the results from this study may be constrained by the particular conditions of the study from where data were collected (for example, sample size, task type, time constraints, and so on), we consider that the use of unimodal and multimodal approaches should be further investigated for practical applications in the wild. In this sense, we acknowledge that our results may not be generalizable to other conditions such as exploratory search tasks or the use of alternative user interfaces (for example, touch, gesture, and brain interfaces). Nevertheless, this variety of scenarios sets the bases for our future work.

Acknowledgments

The work presented in this article was partially funded by Proyecto DICYT 061519GI, Vicerrectoría de Investigación, Desarrollo e Innovación, Universidad de Santiago de Chile and NSF Grant IIS-1717488.

References

Arapakis, I., Jose, J.M., & Gray, P.D. (2008). Affective feedback: an investigation into the role of emotions in the information seeking process. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 395–402). New York: ACM.

Arapakis, I., Konstas, I., Jose, J.M., & Kompatsiaris, I. (2009). Modeling facial expressions and peripheral physiological signals to predict topical relevance. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 728-729). New York: ACM.

Baeza-Yates, R., & Galleguillos, C. (2005). Análisis de consultas a un buscador de la Web chilena. Revista Facultad de Ingeniería-Universidad de Tarapacá, 13(1), 21–29.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In Proceedings of the Seventh International Conference on World Wide Web 7, WWW7 (pp. 107–117). Amsterdam, The Netherlands: Elsevier Science Publishers B. V.

Foresee (2015). ACSI retail report 2014 download American customer satisfaction index. Retrieved from http://www.theacsi.org/news-and-resources/customer-satisfaction-reports/reports-2014/acsi-retail-report-2014/acsi-retail-report-2014-download.

Fox, S., Karnawat, K., Mydland, M., Dumais, S., & White, T. (2005). Evaluating implicit measures to improve web search. ACM Transactions on Information Systems, 23(2), 147–168.

González-Ibáñez, R., Escobar-Macaya, M., & Manriquez, M. (2016). Using low-cost electroencephalography (EEG) sensor to identify perceived relevance on web search. Proceedings of the Association for Information Science and Technology, 53(1), 1–5.

González-Ibáñez, R., & Shah, C. (2011). Coagmento: A system for supporting collaborative information seeking. Proceedings of the Association for Information Science and Technology, 48(1), 1–4.

González-Ibáñez, R. & Shah, C. (2014). Performance effects of positive and negative affective states in a collaborative information seeking task. In CYTED-RITOS International Workshop on Groupware (pp. 153-168). Cham, Switzerland: Springer.

González-Ibáñez, R., & Shah, C. (2016). Using affective signals as implicit indicators of information relevance and information processing strategies. Proceedings of the Association for Information Science and Technology, 53(1), 1–10.

Google (2012). Search quality rating guidelines. Retrieved from https://support.google.com/websearch/answer/179386?hl=es.

Guo, Q. & Agichtein, E. (2012). Beyond dwell time: Estimating document relevance from cursor movements and other post-click searcher behavior. In Proceedings of the 21st International Conference on World Wide Web, WWW '12, (pp. 569–578). New York: ACM.

Gwizdka, J. & Zhang, Y. (2015). Differences in eye-tracking measures between visits and revisits to relevant and irrelevant Webpages. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 811–814). New York: ACM.

Hardoon, D.R., Shawe-Taylor, J., Ajanki, A., PuolamÃ, K., & Kaski, S. (2007). Information retrieval by inferring implicit queries from eye movements. In Artificial Intelligence and Statistics (pp. 179–186).

Jung, S., Herlocker, J.L., & Webster, J.G. (2007). Click data as implicit relevance feedback in web search. Information Processing & Management, 43(3), 791–807.

Lopatovska, I. (2011). Emotional correlates of information retrieval behaviors. In Affective Computational Intelligence (WACI), 2011 I.E. Workshop on (pp. 1–7). IEEE.

Marchionini, G. (2008). Human–information interaction research and development. Library & Information Science Research, 30(3), 165–174.

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, SIGIR '13, (pp. 133–142). New York: ACM.

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).

Saracevic T. (1996). Relevance reconsidered. Information science: Integration in perspectives. In Proceedings of the Second Conference on Conceptions of Library and Information Science, (pp. 201–218), Copenhagen, Denmark.

Sebe, N., Lew, M.S., Sun, Y., Cohen, I., Gevers, T., & Huang, T.S. (2007). Authentic facial expression analysis. Image and Vision Computing, 25(12), 1856–1863.

Teevan, J., Dumais, S.T., & Horvitz, E. (2010). Potential for personalization. ACM Transactions on Computer-Human Interaction (TOCHI), 17(1), 4.

Wang, P., Hawk, W.B., & Tenopir, C. (2000). Users' interaction with world wide web resources: An exploratory study using a holistic approach. Information Processing & Management, 36(2), 229–251.

Warner, D. & Myer, M. (2003). Implicit rating of retrieved information in an information search system. WO Patent App. PCT/US2001/011,959.

White, R.W. & Huang, J. (2010). Assessing the scenic route: measuring the value of search trails in web logs. In Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 587–594). New York: ACM.

DOI: 10.1002/asi