Using Facial Expressions and Peripheral Physiological Signals as Implicit Indicators of Topical Relevance

Ioannis Arapakis, Ioannis Konstas, Joemon M. Jose
Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ
{arapakis,konstas,jj}@dcs.gla.ac.uk

ABSTRACT

Multimedia search systems face a number of challenges, emanating mainly from the semantic gap problem. Implicit feedback is considered a useful technique in addressing many of the semantic-related issues. By analysing implicit feedback information search systems can tailor the search criteria to address more effectively users' information needs. In this paper we examine whether we could employ affective feedback as an implicit source of evidence, through the aggregation of information from various sensory channels. These channels range between facial expressions to neurophysiological signals and are regarded as indicative of the user's affective states. The end-goal is to model user affective responses and predict with reasonable accuracy the topical relevance of information items without the help of explicit judgements. For modelling relevance we extract a set of features from the acquired signals and apply different classification techniques, such as Support Vector Machines and K-Nearest Neighbours. The results of our evaluation suggest that the prediction of topical relevance, using the above approach, is feasible and, to a certain extent, implicit feedback models can benefit from incorporating such affective features.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Relevance Feedback, Search Process*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces; I.5.1 [**Computing Methodologies**]: Pattern Recognition—*Models*

General Terms

Experimentation, Human Factors, Performance

Keywords

Affective feedback, facial expression analysis, physiological signal processing, multimedia retrieval, classification, pattern recognition, support vector machines

MM'09, October 19–24, 2009, Beijing, China.

1. INTRODUCTION

The main challenge multimedia search systems face nowadays emanates from the semantic gap: the semantic difference between a user's query representation and the internal representation of an information item in a collection [47]. Effective search techniques are needed to deal with a variety of multimedia data. Though progress has been made, the effectiveness of existing systems is still limited. The gap is further widened when the user is driven by an ill-defined information need, often the result of an anomaly in his/her current state of knowledge [7]. The formulated search queries, which are used by the search system to locate potentially relevant items, produce results that do not address the users' needs.

To deal with information need uncertainty search systems have employed a range of relevance feedback techniques, which vary from explicit [23, 38] to implicit [1, 5]. Relevance is a key notion of information science, which is intertwined with many interactive processes, such as the act of communication, information seeking, information assessment, reflection, etc. People very often apply relevance assessment when performing some kind of information processing task in order to determine the degree of appropriateness or relation of the perused information item [42]. The value of relevance assessments lies in the progressive disambiguation of the user's information need, which is achieved through an interactive and iterative process known as the relevance feedback cycle. It is suggested that relevance techniques can be utilised for affective retrieval [50, 20].

By analysing explicit and implicit feedback information, search systems can determine topical relevance with better accuracy, offer improved query reformulations and, furthermore, tailor the search criteria to the user needs. However, as discussed in [3], existing feedback techniques determine content relevance only with respect to the cognitive and situational levels of interaction, failing to acknowledge the importance of intentions, motivations and feelings in cognition and decision-making [14, 34]. There is evidence that people naturally express emotion to machines and introduce a wide range of social norms and learned behaviours that guide their interactions with, and attitudes toward, interactive systems and information items [35, 52, 37].

With respect to online search behaviour a number of studies from the field of Library and Information Science (LIS) have provided evidence which reveal that affect can influ-

ence several aspects of the search process, such as the search strategies [30], performance [54, 28] and satisfaction [29]. Positive and negative emotions have been associated with satisfactory search results [49], successful search completion [8] and interest in the process and documents [24, 25]. According to McKechnie and Ross [26], affective variables can play an important role in reading-related information behaviour, especially in the domain of everyday life. Information processing, which occurs during the appraisal process of a goal, an event, or an item, can result in a series of changes in the user's cognitive and affective states [43].

We argue that such changes are often expressed through a psycho-physiological mobilisation that is reflected by a series of more or less observable cues, such as facial expressions, body movements, localised changes in the electrodermal activity, variations in the skin temperature, and many more. Since the significance of an event or information item can vary from low to high, depending on the number of goals or needs that are affected by it, so do these peripheral physiological symptoms vary in intensity and duration.

The modelling and integration of such affective features in the feedback cycle could allow search systems to facilitate a more natural and meaningful interaction. Furthermore, it could improve the quality of the query suggestions, and, potentially, influence other facets of the information seeking process, such as indexing, ranking and recommendation. Eventually, it may be that relevance inferences obtained from this kind of models will also provide a more robust and individualised form of feedback, which will allow us to deal effectively with the semantic gap.

In this paper we explore the role of affective feedback in designing multimedia search systems. We investigate whether we can deduce topical relevance by measuring key physiological signals taken from the user. Our key assumption is that relevance information that derives from the selected sensory channels is correlated to user affective behaviour. However, we do not assume anything about the details of the relationship between users' affective responses and topical relevance; we systematically build our ground truth from the data, using classification and pattern recognition methods.

1.1 Research Questions

The major goal of this study was to investigate the affective response patterns that emerge during the evaluation of different media types (documents, videos), in the context of typical search tasks. The affective behaviour was analysed separately for facial expressions, as well as other peripheral physiological signals. Overall, we examined the following research hypothesis:

***H*$_1$:** Users' affective responses, as determined from automatic facial expression analysis, will vary across the relevance of perused information items.

***H*$_2$:** Users' affective responses, as determined from peripheral physiological signal processing, will vary across the relevance of perused information items.

The rest of the paper is structured as follows: Section §2 reviews existing work on the use of eye-tracking to predict of topical relevance, the employment of affective feedback for the improvement of recommendations and the application of affective summarisation. Section §3 presents the experimental design of our study and its sub-components. Section §4 discusses the models and the classification techniques that were employed. Section §5 provides details about the preprocessing and feature engineering that were applied on our data set. Section §6 presents the implications of the results on the prediction of topical relevance with the use of affective feedback, while section §7 describes briefly the results our follow-up study. Finally, Section §8 concludes the paper and Section §9 provides future directions.

2. RELATED WORK

To our knowledge, the prediction of topical relevance and user interest through the modelling of affective behaviour is a new and unexplored area. However, examples of earlier work exist in several similar fields. In [41, 36] the authors adopt a user-centred approach and develop models that can infer relevance implicitly from eye-movement data. In addition, they combine successfully the measured implicit feedback with a database of user preferences, offering evidence that collaborative filtering and implicit feedback can be used individually, or to complement standard textual content-based filtering.

In [4], the authors present a novel video search environment that aggregates information from users' affective behaviour, by applying real-time facial expression analysis. The affective information is used to determine the topical relevance of perused videos and enrich the user profiles. The combination of different modalities (facial expression data, interaction data, etc.), along with the integration of affective features, allowed the facilitation of meaningful recommendations of unseen videos.

Mooney et al. [27], performed a preliminary study of the role of physiological states, in an attempt to improve data indexing for search and within the search process itself. Users' physiological responses to emotional stimuli were recorded using a range of metrics (galvanic skin response, skin temperature, etc.). The study provides some initial evidence that support the use of physiological signals in that context.

A content-centred approach has been adopted in studies [10] and [19]. In [10], a set of affective features was extracted from audio content, and was annotated using a set of labels with predetermined affective semantics. The audio features, which consisted of speech, music, special effects and silence, were analyzed in terms of the affective dimensions of arousal and valence. These measurements were combined to form a plot, known as the *affect curve*. Similarly, in [19] the authors model video content using a selection of low level audio (signal energy, speech rate, inflection, rhythm duration, voice quality) and visual features (motion). Again, this framework is based on the dimensional approach to affect, with the video content represented as a set of points in a two-dimensional affect space, of arousal and valence, that reliably depict the expected emotional transitions across the video (as perceived by a viewer).

Finally, in [48] the authors propose an approach to affective ranking of movie scenes which is based on viewers' affective responses, as well as content-based multimedia features. The latter features appear to capture important aspects of the events that characterise every scene and are apparently correlated with the viewers' self-assessments of arousal and valence. Furthermore, the evidence provided, suggests that peripheral physiological signals can be used to characterise and rank video content, even though the variation of users' affective responses emphasises the need for more personalised emotional profiles.

While the above studies look into different aspects of multimedia retrieval none of them accounts for the affective dimension of search behaviour, in relation to topical relevance, with the exception of [4]. The contributions of this paper include not only a new approach in pattern analysis of affective data, deriving from facial expressions and physiological signals, but also the finding of encouraging classification rates. Regarding facial expression analysis, we also test the performance of models trained on motion-unit data (a category of low-level features), instead of emotion categories, which prove to perform better than the latter. In addition, we developed an annotated collection of affective data that accounts for different types of stimuli (text and video). It is anticipated that a correlation exists between affective behaviour and topical relevance. Our models have been trained on data that derive from varied content and type, making their application applicable to different topics and media, which is a step towards an effective multimedia search.

3. EXPERIMENTAL METHODOLOGY

Even though physiological response patterns and emotion behaviour are observable there are no objective methods of measuring the subjective experience [44]. Very often, the emotional experience is captured using a combination of think-aloud protocols and forced-choice or free-response reports. In some cases it is decomposed and examined through the application of a multi-modal analysis, using a range of sensory data [33]. Two of the most popular approaches in affective behaviour analysis have been the discrete-categories and the dimensional approach.

Discrete emotion theorists suggest the existence of six or more basic emotions (happiness, sadness, anger, fear, disgust and surprise), which are universally displayed and recognised [18]. The term "basic" is primarily used to denote elements that can be combined to form more complex or compound emotions. However, emotions can be often experienced as mixed or blended, making the classification into a limited number of categories too restrictive. The existence of basic emotions is supported by evidence of cross-cultural universals for facial expressions and antecedent events, as well as the presence of such states in other primates [17].

The dimensional approach suggests the characterisation of emotions in terms of a two-dimensional affect space of arousal and valence [40]. Valence represents the pleasantness of the stimuli along a bipolar continuum, between a positive and a negative pole, while arousal indicates the intensity of the emotion [40]. This dimensional taxonomy of emotions treats all emotion categories as varying quantitatively from one another and represents their relationships as distances within the affect space.

In this study we employed eMotion, an automatic facial expression recognition system [51] that applies the first approach. Recent research indicates that emotions are primarily communicated through facial expressions [31] and provide useful cues (smiles, chuckles, smirks, frowns, etc.) that are considered an essential aspect of our social interaction. We, furthermore, recorded a range of peripheral physiological signals using a set of wearable devices (see section §3.2), which gave us a more fine-grained imprint of subjects' affective states. Our motivation for using the above tools was to establish a relation between facial expressions, and other sensory data collected, to the users' affective responses towards topical relevance.

3.1 Design

This study used a repeated-measures design. There were three independent variables: media type (with two levels: "document" and "video"), topical relevance (with two levels: "relevant" and "irrelevant") and classifier (with two levels: "SVM" and "KNN"). The media type levels were controlled by assigning topics associated with a document or a video collection, accordingly. Topical relevance levels were controlled by presenting results that belonged exclusively to one of the two categories, using the ground-truth associated with the collection (TREC's qrel) as a selection criterion. The classifier levels controlled by applying a different classification method. The dependent variables were: (i) accuracy, (ii) precision, (iii) recall (iv) emotional responses (as determined from facial expressions) with respect to topical relevance, (v) emotional responses (as determined from physiological signals) with respect to topical relevance.

3.2 Apparatus

For our experiment we used two desktop computers, equipped with conventional keyboards and mouse. The first computer (server) hosted the retrieval system (Verge Engine) [53], the two test collections (document & video), the SQL database that held the interaction data and eMotion. The second computer (client) provided access to the GUI environment of Verge Engine. It also logged subjects' desktop actions, starting, finishing and elapsed times for interactions, and click-throughs, using a custom-made script. The script was executed in the background and stored the above information to the server's database.

In addition, we installed a web-camera on the setting: a Live! Cam Optia AF web camera, with a 2.0 megapixels sensor. The camera was used for recording the subjects' expressions, as well as apply real-time facial expression analysis. Finally, we used two unobtrusive wearable devices to capture subjects' physiological signals: (i) Polar RS800 Heart Rate Monitor [21], and (ii) BodyMedia SenseWear® Pro3 Armband [2]. All devices and systems were logging using a common system time.

3.2.1 Test Collections & Search Tasks

For the video indexing we used TRECVID 2007 test collection, which contains 200 hours of videos. The collection was built using the Dutch television archive and covers a range of video genres, such as news magazine, science news, news reports, documentaries, educational programming, and other [45]. For the document indexing we used TREC 9 (2000) Web Track, which is a 1.69 billion document subset of the VLC2 collection of 10 gigabyte size [6]. By introducing both video and textual information we allowed for the collection and study of affective feedback in response to different sources of media stimuli, thus developing a media independent feedback mechanism.

In this work, we retained the original content of the TREC topics, which in terms of delimiting the area of searching appear to be effective enough. The basic assumption behind the topic frame is that an information need should be treated as static and accurately defined, providing an objective measure for precision-recall. For each media type category we selected four different TREC topics of varying content and type. The subjects had the option, for each media type, to perform two tasks of their choice. For each topic we had pre-selected ten relevant and ten irrelevant results, using the

ground-truth associated with the collection as our selection criterion. However, considering the variability of personal relevance judgements, some relevant/irrelevant items might have not been perceived as such. We, therefore, drew our final decision on the class (relevant, irrelevant) of each item based on the subjects' explicit judgements.

3.2.2 Search Interface

For the completion of the search tasks we used a customised version of Verge Engine [53], which worked on top of TRECVID 2007 and TREC 9 (2000) Web Track test collections. Verge Engine was selected for its versatile interactive environment, which combines several basic retrieval functionalities (visual similarity search module, textual information processing module, etc.) with a user-friendly interface, that can support adequately the submission of search queries and the retrieval of results in the TREC standard format. In our modified version of Verge Engine, each result was represented by a link, accompanied by a short summary (in the case of documents), or a thumbnail, along with some meta-information (in the case of videos). The meta-information consisted of the shot-id, associated keywords and the duration of the video.

The evaluated version applied a layered architecture approach. The first layer was responsible for supporting any interaction occurring during the early stages of the search process (such as query formulation and search execution). Any output generated during that phase was presented in the second layer. From there, the subjects could select and preview any of the retrieved items. The content of an item was shown in a separate panel in the foreground, which constitutes the third layer of our system.

The main purpose of the layered architecture was to isolate the viewed content from all possible distractions that reside on the desktop screen; therefore, establishing additional ground truth that allowed us to relate users' emotional responses to the source of stimuli (in our case, the perused information items). This was an important aspect of our experimental methodology, since we were interested in isolating content-particular user emotions. Upon viewing the video or document, the subjects had to evaluate: (i) the degree of relevance, and (ii) the emotional impact of the viewed content.

3.2.3 Questionnaires

The subjects completed an Entry Questionnaire at the beginning of the study, which gathered background and demographic information, and, furthermore, inquired about previous experience with multimedia and online searching. The information obtained from it was used to characterise the subjects, but not in subsequent analysis. A Post-Search Questionnaire was also administered at the end of each task, to elicit subjects viewpoint on certain aspects of the search process. The questions were divided into four sections that covered the search session, the encountered task, the emotional experience (with respect to the viewed content) and the encountered results.

All of the questions included in the questionnaire were forced-choice type, with the exception of a single question that requested a written description. This description asked for the event that elicited the emotional episode, in addition to details regarding what took place and the consequences it had for the participant. Finally, an Exit Questionnaire was introduced at the end of the study. The questionnaire gathered information on the topic descriptions and the perceived relevance criteria, as well as subjects' views of the importance of affective feedback, with respect to usability and ethical issues.

3.2.4 Facial Expression Recognition System

Facial expressions have been associated in the past with universally distinguished emotions, such as happiness, sadness, anger, fear, disgust, and surprise [16]. Recent research also indicates that emotions are primarily communicated through facial expressions [31], which provide useful cues (smiles, chuckles, smirks, frowns, etc.) that are considered an essential part of our social interaction [39].

In this study we applied real-time facial expression analysis using the system mentioned in section §3.2. The process occurred as follows: initially, eMotion would detect certain facial landmark features (such as eyebrows, the corners of the mouth, etc.) and construct a 3-dimentional wireframe model of the face, consisting of a number of surface patches wrapped around it. After the construction of the model, head motion or any other facial deformation would be tracked and measured in terms of motion-units (MU's), and, finally, classified into one of the seven detectable emotion categories.

Even though eMotion follows the categorical approach we did not employ categorical data for the training of our models. Instead, we used the MU's data, which is a low-level category of features very similar to Ekman's action-units (AU's). MU's measure the intensity of an emotion indirectly, by tracking the presence and degree of changes in all facial regions associated with it. Moreover, MU's allowed us to associate the captured facial expressions with a wider range of affective and cognitive states, which are not accounted for during the meta-classification that eMotion applies.

Automatic systems are an alternative approach to facial expression analysis [33] and have exhibited performance which is comparable to that of trained human recognition (87%). eMotion applies a generic classifier that has been trained on a diverse data set, combining data from the Cohn-Kanade database. Its main advantage is its reasonable performance across all individuals, irrespectively of the variation introduced from mixed-ethnicity groups. Results of the person-dependent and person-independent tests presented in [51] support our performance-related assumptions.

3.2.5 Biometrics

Emotions can be expressed through several sensory channels and are reflected by a series of more or less observable cues, such as localised changes in the electrodermal activity, variations in the skin temperature, and many more. It has been shown that transitions between emotional states are correlated with temporal changes in physiological states [12], which cannot be easily faked. In this work we monitor the subjects affective responces by employing a series of physiological signals, such as heart-rate, galvanic skin response and skin temperature. These modalities have been used to measure negative emotions, such as stress and anxiety [55], and user interest towards multimedia content [46, 48] and games [11].

For recording the subjects' physiological signals we used Polar RS800 and BodyMedia SenseWear ® Pro_3 Armband. Polar RS800 Heart Rate Monitor consists of Polar RS800

Running Computer, a wrist-watch that displays and records the heart rate data, an elastic strap with two electrodes and Polar WearLink® 31, a wireless transmitter. The elastic strap is worn around the chest (bellow the chest muscles), allowing the built-in soft textile electrodes to detect the heartbeat and then transmit the heart rate signal to the running computer, via the Polar WearLink® 31, which is attached to the strap.

BodyMedia SenseWear ® Pro_3 Armband is an unobtrusive, lightweight, multi-sensor hub, which is worn above the tricep area of the right arm. It can measure simultaneously five low-level physiological key metrics, namely: (i) galvanic skin response, (ii) skin temperature, (iii) near-body ambient temperature, (iv) heat flux, and (v) motion, via a 3-axis accelerometer. From those vital sign streams it can produce accurate statements about the human body states and behaviours. Moreover, the existence of multiple sensors allows for the disambiguation of contexts, which a single sensor would have not interpreted accurately.

3.3 Participants

Twenty-four healthy subjects, all employees at the Institute of Telematics and Informatics (ITI), applied for the study through an organisational-wide ad. The subjects were in their majority of Hellenic nationality and had a mixed-educational background (5 with PhD degrees, 8 MSc degrees, 10 with BSc degrees and 1 other). They were all proficient with the English language (4 native, 13 advanced, 5 intermediate and 2 beginner speakers).

Out of 24, 14 were male and 10 were female and were between 23-38 years of age (M=28.83 SD=4.13). They had an average of 8.82 years of online searching experience and all claimed to have been using at least one search service in the past (with the most popular being "Google Video" and "Youtube"). On average, the subjects reported dealing with videos, photographs and images once or twice a day (M=5.16 SD=1) and carrying out image or video searches once or twice a week (M=4.29 SD=1.08). The frequency was measured using a 6-point scale (1="Never", 2="Once or twice a year", 3="Once or twice a month", 4="Once or twice a week", 5="Once or twice a day", 6="More often").

3.4 Procedure

The user study was carried out in the following manner. The formal meeting with the subjects took place in the laboratory setting. At the beginning of every session the subjects were given an information sheet, which explained in detail the conditions of the experiment. They were then asked to sign a Consent Form and were also notified about their right to withdraw at any point during the study, without having their legal rights or benefits affected. Finally, they were given an Entry Questionnaire to fill in. The session proceeded with a a brief tutorial on the use of the search environment, followed by a calibration of the sensory devices (Polar RS800 & BodyMedia SenseWear ® Pro_3 Armband.) and the cameras. To ensure that the subjects' faces would be visible to the camera at all times we encouraged them to keep a proper posture, by indicating health and safety measures.

Each subject completed four search tasks in total, two for each media type. In every task they were handed four topics and were asked to proceed with the one they considered more interesting. For each topic the subjects were asked to evaluate ten pre-selected results (either exclusively relevant or irrelevant), under the assumption that these were retrieved by the search system for the given topic, while considering the relevance criteria provided by the scenario description. To negate the order effects we counterbalanced the task distribution by using a Latin Squares design. The subjects were asked every time to evaluate as many (if not all) results as possible and were given 10 minutes to complete their task, during which they were left unattended to work.

At the end of each task, the subjects were asked to complete a Post-Search Questionnaire. An Exit Questionnaire was, additionally, administered at the end of each session. The subjects were encouraged to ask questions and were, once again, informed about their right to withdraw and have any data gathered on them instantly and permanently destroyed.

4. MODELS

We explore the role of affective feedback in designing multimedia search systems. The modelling goal is to promote a more natural and meaningful interaction by predicting with reasonable accuracy topical relevance of videos and online documents. We employ sensory data that derive from facial expressions and other peripheral physiological signals as the only feedback information. From the latter signals we extract a set of features, and performed discriminant analysis, using a range of classification methods, such as Support Vector Machines (SVM) and K-Nearest Neighbours (KNN) clustering. We do not assume anything about the relationship between these features (which we consider indicative of users' affective behaviour) and topical relevance, but rather follow a straightforward classification approach, using the ground truth that is associated with our training data.

4.1 Support Vector Machines

We use libSVM[1], an implementation of support vector machines (SVM), to predict the category (two classes: relevant or irrelevant) of documents and videos which were viewed by the users. Our approach utilises an efficient method that can deal with a difficult, multi-dimensional classification problem. We trained our models using a radial basis function (RBF) kernel, which, among the basic four SVM kernels (linear, polynomial, radial basis function, sigmoid), was considered as a reasonable first choice. Moreover, the RBF kernel is preferable, since it encounters less numerical difficulties and has a limited number of hyper-parameters.

To optimise the performance of SVM model we performed a grid-search on the parameters C (cost) and γ (gamma), using cross-validation during which we tried exponentially growing sequences of C and γ. However, since performing a full grid-search can be time consuming, we initially used a coarse grid and then, after identifying a "good" region, we performed a finer grid search on that region. The end-purpose was to identify the optimal set of (C,γ) so that every classifier we trained could achieve the best possible (tuning-wise) accuracy score on in testing data.

Independently of the kernel function and the parameters, we trained three different categories of SVM models: (i) plain models, using different training and test sets, (ii) models using 5-fold cross-validation, to compensate for the small

[1]http://www.csie.ntu.edu.tw/c̃jlin/libsvm/

Facial expression features (2-dimensional camera)
Emotion categories
Neutral
Happiness
Surprise
Anger
Disgust
Fear
Sadness
Motion Units
Motion Vector 1-10
Peripheral physiological metrics
Transverse Acceleration Point
Longitudinal Acceleration Point
Average Heat Flux
Average Skin Temperature
Average Transverse Acceleration
Average Longitudinal Acceleration
Average Near-Body Temperature
Transverse Acceleration MAD
Longitudinal Acceleration MAD
Average GSR

Table 1: Features used to represent user affective behaviour (in terms of topical relevance)

size of the training/test sets, and (iii) two-layer hierarchical SVM models, with five (5WC) and ten weak (10WC) classifiers. The last category of models uses n weak classifiers, each trained on a different subset of the training set. The whole training set is then predicted once, and the output of each weak classifier is used to train the meta-classifier. This hierarchical framework improved the accuracy of classification, in all cases, by a small percentage.

4.2 K-Nearest Neighbours

We used the implementation of IB1 from the Memory-Based Learning platform TiMBL [13], which uses non-trivial data structures and speed-up optimisations superior to other implementations (e.g. WEKA[2]) of KNN algorithms. In the training phase TiMBL applies a discount function to each feature based on their Gain Ratio. In the test phase it offers the opportunity to choose from a range of metrics to influence the definition of similarity of neighbours. In order to fine tune the performance of the KNN model we used all combinations of these similarity metrics, namely weighted overlap, modified value difference metric, Jeffrey divergence and Levenshtein distance. We also iterated over the number of nearest neighbours with $k = 1$ to 15.

5. DATA ANALYSIS

Out of the 963 browsing instances that took place in this study, 461 correspond to documents and 502 to videos. From these we collected data on: (i) 474 viewing sessions (285 relevant videos, 189 irrelevant videos), for the video category, and (ii) 429 reading sessions (138 relevant documents, 291 irrelevant documents), for the document category. Overall, for the facial expressions category (2-dimensional camera), we accumulated 286564 feature vectors (194798 for document sessions, 91766 for video sessions), while for the biometrics category we gathered a considerably smaller set of 26861 feature vectors (18098 for document sessions, 8763 for video sessions). The data gathered from the heart monitor are excluded from this analysis, since the preliminary analysis did not reveal any information gain.

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

5.1 Features

One of our main objective was to develop a sufficiently rich set of features that will allow us to determine whether a user perceives an information item (document, video) as relevant to his/her need. This information was obtained using the instruments described in section §3.2.4 and §3.2.5. The original number of features that we acquired summed to 53. However, out of the 53, we eventually applied in our analysis only 29, after performing feature selection. By doing so we concluded to a subset of relevant features that allowed for the building of robust learning models, while introducing minimum noise. The reduction of the feature set also reduces the dimentionality of the problem and, thus, the time required to apply the learning algorithms.

The features we use to model user affective behaviour are summarised in Table 1. For clarity, we categorise the features into two groups: (i) facial expression features, and (ii) peripheral physiological features. Due to the different sample rate of each tool, the output of each sensory channel is not synchronised with the rest. However, this did not pose a problem in our analysis since it was not a necessary step in the training process, as long as there were enough training data for each reading/viewing session. Finally, the analysis was performed on a frame-basis.

Facial expression features: In summary, the user affective behaviour is represented by a selection of features that have directly measured values. Most of these attributes have been associated in the past with important affective and cognitive processes. A well-known study by Paul Ekman [15] has shown that certain universal facial expressions, when spontaneously displayed, signal emotions of anger, disgust, fear, happiness, and surprise. Facial expressions are generally regarded as essential aspects of human social interaction. The face provides conversational signals, which do not only clarify our current focus of attention [32] but also regulate our interactions with the surrounding environment and the organisms that inhabit it.

Peripheral physiological signals: Physiological signals, similarly to facial expression recognition, can play a significant role in emotion recognition. However, the physiological responses among individuals is expected to be more diverse and, therefore, it is generally harder to determine whether these transitions occur due to a change in the affect state or other factors, e.g. cognitive processes, sensory stimuli, etc. In this study, we used a range of sensory input, such as skin temperature, galvanic skin response, etc., which can be measured easily and unobtrusively, and are considered reliable affect-specific characteristics.

Galvanic skin response represents the activity of the autonomic nervous system [9]. Changes in the electrical properties of the skin, due to the activity of the sweat glands, is physically interpreted as conductance. The sweat glands, which are distributed all over the skin, receive input from the sympathetic nervous system, making it a good indicator of the level of emotional arousal due to external sensory or cognitive stimuli. Variations in the skin temperature are mainly

Model	Accuracy (%)	Precision (%)	Recall (%)
Baseline (Random selection)	50.0	0.0	0.0
Videos			
*SVM Training & Test Set**	63.9	63.7	67.0
*SVM (5WC)**	64.9	64.5	68.8
*SVM (10WC)**	64.4	64.7	69.5
*KNN Training & Test Set (Levenshtein, K=15)**	56.9	58.6	51.5
Documents			
*SVM Training & Test Set**	57.1	60.8	52.9
*SVM (5WC)**	57.6	60.5	56.5
SVM (10WC)	57.9	60.6	57.7
*KNN Training & Test Set (Levenshtein, K=15)**	51.7	49.0	51.4

Table 2: Results for models trained on facial expressions (motion units).

Model	Accuracy (%)	Precision (%)	Recall (%)
Baseline (Random selection)	50.0	0.0	0.0
Videos			
*SVM Cross-Validation (5 folds)**	66.5	66.2	67.6
*KNN Cross-Validation (5 folds, Jeffrey Divergence, K=1)**	63.7	65.5	64.4
Documents			
*SVM Cross-Validation (5 folds)**	60.4	61.9	54.1
*KNN Cross-Validation (5 folds, Levenshtein, K=15)**	55.6	49.3	54.3

Table 3: Results for models trained on peripheral physiological signals.

the result of localised changes in the blood flow, caused by vascular resistance or arterial blood pressure [22]. Skin temperature, which reflects the autonomic nervous system activity, is also another reliable indicator of the underlying affect state.

5.2 Preprocessing

For both categories of sensory information, as well as all media types, we randomly separated the data into two sets (training and test), using an equal number of documents. This resulted in two sets with approximately the same number of feature vectors. By balancing the training and test sets we prevented over-fitting and, additionally, compensated for the originally uneven size of the data sets. To our knowledge, none of the instruments we used pre-processes the data. We, therefore, had to scale them before applying any classification method, to avoid having attributes in greater numeric ranges dominating those in smaller numeric ranges. We, finally, performed feature selection, using as a selection criterion the information gain of each feature and concluded to a representative set of features, which was used to train a model with the best feature-wise discriminating ability.

6. RESULTS

In this section we present the experimental findings of our study, based on 96 search sessions that were carried out by 24 subjects. Out of the many results, we are reporting those that refer to our models and present only the questionnaire data that refer to ethical and usability issues of our system, due to limited space. We measured the performance of all models using the standard information retrieval metrics of accuracy, precision and recall. Accuracy was computed as the fraction of items in the test set for which the models' predictions were correct.

6.1 Models

For each classification method we present only the model which achieved the best performance among the rest in its category. The results are shown in Tables 2 and 3. The McNemar's test was applied to check for statistically significant variation against the baseline. Models marked with (*) got significantly different results, compared to the baseline model, with $p < 0.005$. The baseline represents random choice and, since the class of an information item (document, video) can be either relevant or irrelevant, it is set to 50%. Among all the models, the SVM held the best performance, giving a reasonable, though rather noisy, prediction of topical relevance. The boosting that we performed, using either 5 or 10 weak classifiers, gave a slight increase in the accuracy. The model that held the best accuracy, from those trained on facial expressions, was the SVM with 5 weak classifiers (64.9%), followed by the SVM with 10 weak classifiers (57.9%), for the documents category. From the models trained on peripheral physiological signals the SVM with 5-fold cross-validation held the best accuracy for both video (66.5%) and documents (60.4%).

With the exception of one SVM model, all the rest had a statistically significant difference in their classification accuracy, compared to the baseline model. The discriminative KNN model had in all cases the lowest performance. This was an anticipated outcome, since the KNN algorithm cannot deal with the same efficiency the potential non-linearity of the feature space, as in the case of the SVM, which exhibits a more sophisticated discriminative ability. The models that were trained on the *emotion categories* of facial expression data produced low accuracy, unlike the models trained on the MU's, and were omitted from further discussion. We speculate that this was a result of the meta-classification that eMotion applies when categorising each captured expression into one of the seven recognisable emo-

tion categories. Another potential reason for the low accuracy is the reduction of the dimentionality of the feature space, which is a result of the meta-classification applied.

6.2 Questionnaires

A 5-point scale Likert scale was used in all questionnaires. Questions that ask for user rating on a unipolar dimension have the positive concept corresponding to the value of 1 (on a scale of 1-5) and the negative concept corresponding to the value of 5. Questions that ask for user rating on a scale of 1-5 represent in our analysis stronger perception with high scores and weaker perception with low scores. When asked about the importance (M=3.2917, SD=1.1221) and helpfulness (M=3.3333, SD=1.0072) of a search system that integrates affect aware technologies, the subjects did not exhibit a major trend in their views. The same applies for their view in terms of privacy and intrusiveness (M=3.1250, SD=1.1156), which is a positive outcome, considering that the success of such a system depends highly on its acceptability. One interesting finding is that, overall, users' do not perceive as unethical to have their emotional behaviour monitored (M=2.5000, SD=1.1034). Perhaps this is one aspect of human-computer interaction that imitates human-human interaction and, therefore, feels more tolerable and natural. Finally, the subjects consider an affect-aware search system better than existing search tools that do not integrate similar emotion-detection modules (M=3.5000, SD=0.7223).

7. FOLLOW-UP STUDY

To further evaluate our models we performed a follow-up study, using six researchers from our institute who volunteered for the evaluation. The study was based on realistic re-assessments of the documents, from the same test collection and topics as in the main experiment. The same set of features was also recorded and used to test our models. Overall, we collected biometric data for 29 video sessions, and facial expression data for 69 document sessions and 68 video sessions. The results of the evaluation revealed that, even under realistic conditions, most of our models can attain a performance which is better than random. However, we only regard this evidence as a positive indication of the models' predictive capabilities, rather than the ground truth. A more extensive evaluation needs to be conducted, using a sufficiently larger test set.

8. DISCUSSION & CONCLUSIONS

We have presented a controlled experimental framework where the subjects evaluated the relevance of videos and documents, according to the topic at hand. The purpose of the study was to analyse quantitatively the affective responses of the users, while they were evaluating the results retrieved by a search engine in an attempt to locate relevant items. Our experiment was designed to resemble actual search scenarios. We took measurements of facial expressions and key physiological signals and we used classification techniques (SVM & KNN) to train models capable of discriminating successfully the category (relevant vs irrelevant) of the information items.

One facet of affect recognition is developed here for the first time: classification of user affective responses from facial expression and physiological data, gathered from many subjects. The accumulated evidence supports both of our hypotheses, namely that users' affective responses, as determined from the observation of their facial expressions and other peripheral physiological signals, will vary across the relevance of perused information items. Our best performing model attained accuracy of 66.5%, which is reasonably better than the baseline. We are aware that in some cases the content itself might have induced emotional reactions which were unrelated to topical relevance. Nevertheless, we feel that we can deal with this issue in the future by introducing a multi-modal framework for affective feedback.

Overall, the performance of our models for the video topics indicates that audio-visual content is a stronger stimuli, compared to textual information which induces milder affective responses. This suggests that multimedia retrieval might prove a more suitable area of application for affective feedback. Both categories of sensory data (facial expressions, physiological signals) proved to be a good first choice, but alternative modalities should be also explored. With respect to facial expressions, we found that low-level features perform better, compared to high-level information such as emotion categories. One potential explanation is that there is unnecessary redundancy of the values produced by eMotion's meta-classifier output, which cannot be modelled efficiently. In addition, a naturalistic experimental setting that applies hidden recording would be in favour of acquiring spontaneous and authentic facial expressions [3], thus producing data with more discriminative characteristics.

The results of our follow-up evaluation indicate that the application of our models to new users and under a different setting is possible, and to a certain extent models can benefit from taking into account user affective behaviour. Our model-based approach was designed to be as independent as possible from the viewed content and context, therefore, making its application generalizable to a range of different search topics and multimedia. This is perhaps the most significant contribution of this work, since it will potentially influence other aspects of the search process, such as relevance feedback, ranking, recommendation techniques, as well as offer new insight to the semantic gap problem.

In conclusion, this is a new field without much existing work. We feel that the quality of our results is good enough to indicate that affective feedback is a promising area of research and that it can be further developed. Finally, since there are no other systems available for direct comparison, our system holds the best accuracy achieved, so far, in the deduction of topical relevance using affective information. Our study provides the next step towards the modelling of affective feedback, in the context of online information seeking.

9. FUTURE WORK

Apparently, there is still a lot of room for improvement, both with respect to the accuracy of the classifiers, as well as the feature selection. The future goal would be to experiment with additional pattern recognition techniques and modalities, in order to optimise the extraction of affective information from user search behaviour. In addition, we intend to apply information fusion and combine the predictions of models trained on different sensory data (facial expressions, biometrics, EEG), thus improving their discriminative ability. We will also investigate further the effect of personalised data, compared to data deriving from multiple users, on the performance of our models, as well as

examine the differences in users' affective responses during self-generated and assigned tasks. It is possible that users exhibit different affective responses while searching for information relevant to their own tasks and problems. Finally, we are interested in comparing the performance of our classifiers against existing implicit feedback techniques, which account only for interaction data.

10. ACKNOWLEDGMENTS

The research leading to this paper was supported by the European commission, under the contracts FP6-033715 (MIAUCE) and FP6-027122 (SALERO). Our thanks to MKLab for participating in this study.

11. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26. ACM, 2006.

[2] D. Andre, R. Pelletier, J. Farringdon, S. Safier, W. Talbott, R. Stone, N. Vyas, J. Trimble, D. Wolf, S. Vishnubhatla, S. Boehmke, J. Stivoric, and A. Teller. The development of the sensewear® armband, a revolutionary energy assessment device to assess physical activity and lifestyle. Technical report, BodyMedia, Inc., 2006.

[3] I. Arapakis, J. M. Jose, and P. D. Gray. Affective feedback: an investigation into the role of emotions in the information seeking process. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 395–402. ACM, 2008.

[4] I. Arapakis, Y. Moshfeghi, H. Joho, R. Ren, D. Hannah, and J. M. Jose. Enriching user profiling with affective features for the improvement of a multimodal recommender system. In *Conference on Image and Video Retrieval*, 2009.

[5] R. Badi, S. Bae, J. M. Moore, K. Meintanis, A. Zacchi, H. Hsieh, F. Shipman, and C. C. Marshall. Recognizing user interest and document value from reading and organizing activities in document triage. In *Proceedings of the 11th international conference on Intelligent User Interfaces*, pages 218–225, New York, NY, USA, 2006. ACM.

[6] P. Bailey, N. Craswell, and D. Hawking. Engineering a multi-purpose test collection for web retrieval experiments. *Inf. Process. Manage.*, 39(6):853–871, 2003.

[7] N. J. Belkin. Anomalous state of knowledge for information retrieval. *Canadian Journal of Information Science*, 5:133–143, 1980.

[8] D. Bilal and J. Kirby. Differences and similarities in information seeking: children and adults as web users. *Information Processing and Management: an International Journal*, 38(5):649–670, 2002.

[9] W. Boucsein. *Electrodermal activity*. Plenum Press, New York, 1992.

[10] C. H. Chan and G. J. F. Jones. Affect-based indexing and retrieval of films. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 427–430, New York, NY, USA, 2005. ACM.

[11] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun. Boredom, engagement and anxiety as indicators for adaptation to difficulty in games. In *MindTrek '08: Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era*, pages 13–17, New York, NY, USA, 2008. ACM.

[12] R. R. Cornelius. Theoretical approaches to emotion. *Proc ISCA: workshop on Speech and Emotion*, pages 3–11, 2000.

[13] W. Daelemans and A. van den Bosch. *Memory-Based Language Processing (Studies in Natural Language Processing)*. Cambridge University Press, New York, NY, USA, 2005.

[14] A. R. Damasio. *Descartes Error: Emotion, Reason, and the Human Brain*. Putnam/Grosset Press, 1994.

[15] P. Ekman. *Facial Expressions*. Handbook of Cognition and Emotion. John Wiley & Sons Ltd., New York, 1999.

[16] P. Ekman. *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. Times Books, New York, 2003.

[17] P. Ekman. *Unmasking the face*. MA: Malor books, Cambridge, 2003.

[18] P. Ekman and H. Oster. Facial expressions of emotion. *Annual Review of Psychology*, 30(1):527–554, 1979.

[19] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *Multimedia, IEEE Transactions on*, 7(1):143–154, 2005.

[20] F. Hopfgartner. A news video retrieval framework for the study of implicit relevance feedback. In *Proceedings of the Second International Workshop on Semantic Media Adaptation and Personalization*, pages 233–236, Washington, DC, USA, 2007. IEEE Computer Society.

[21] http://www.polarusa.com.

[22] H. Kataoka, H. Kano, H. Yoshida, A. Saijo, M. Yasuda, and M. Osumi. Delelopment of a skin temperature measuring system for non-contact stress evaluation. In *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 2, pages 940–943, 1998.

[23] J. Koenemann and N. J. Belkin. A case for interaction: a study of interactive information retrieval behavior and effectiveness. In *CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 205–212, New York, NY, USA, 1996. ACM.

[24] J. Kracker. Research anxiety and students' perceptions of research: an experiment. part i: Effect of teaching kuhlthau's isp model. *Journal of the American Society for Information Science and Technology*, 53(4):282–294, 2002.

[25] I. Lopatovska and B. Mokros, H. Willingness to pay and experienced utility as measures of affective value of information objects: Users' accounts. *Information Processing and Management: an International Journal*, 44(1):92–104, 2008.

[26] M. Lynne, R. C. Sheldrick, and R. Paulette. *Affective*

dimensions of information seeking in the context of reading. Medford, NJ: Information Today, 2007.

[27] C. Mooney, M. Scully, G. J. Jones, and A. F. Smeaton. *Investigating Biometric Response for Information Retrieval Application*, volume 3936/2006 of *Lecture Notes in Computer Science*. Springer Berlin, 2006.

[28] D. Nahl. Learning the internet and the structure of information behavior. *Journal of the American Society for Information Science*, 49(11):1017–1023, 1998.

[29] D. Nahl. Measuring the affective information environment of web searchers. In *Proceedings of the American Society for Information Science and Technology*, volume 41, pages 191–197, 2004.

[30] D. Nahl and C. Tenopir. Affective and cognitive searching behavior of novice end-users of a full-text database. *Journal of the American Society for Information Science*, 47(4):276–286, 1996.

[31] M. Pantic and L. Rothkrantz. Expert system for automatic analysis of facial expression. *Image and Vision Computing Journal*, 18(11):881–905, July 2000.

[32] M. Pantic and L. J. M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. In *Proceedings of the IEEE*, pages 1370–1390, 2003.

[33] M. Pantic, N. Sebe, J. F. Cohn, and T. Huang. Affective multimodal human-computer interaction. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 669–676, New York, NY, USA, 2005. ACM.

[34] H.-R. Pfister and G. Böhm. The multiplicity of emotions: A framework of emotional functions in decision making. *Judgment and Decision Making*, 3:5–17, 2008.

[35] R. W. Picard, A. Wexelblat, and C. I. N. I. Clifford I. Nass. Future interfaces: social and emotional. In *CHI '02 extended abstracts on Human factors in computing systems*, pages 698–699, 2002.

[36] K. Puolamäki, J. Salojärvi, E. Savia, J. Simola, and S. Kaski. Combining eye movements and collaborative filtering for proactive information retrieval. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 146–153, New York, NY, USA, 2005. ACM.

[37] B. Reeves and C. Nass. *The media equation: how people treat computers, television, and new media like real people and places*. Cambridge University Press, New York, NY, USA, 1996.

[38] Y. Rui and S. Huang, T. Optimizing learning in image retrieval. In *IEEE Proceedings of Conference on Computer Vision*, volume 1, pages 236–243, 2000.

[39] A. Russell, J., J. Bachorowski, and J. Fernandez-Dols. Facial and vocal expressions of emotion. *Annual Review of Psychology*, 2003.

[40] J. A. Russell and J. H. Steiger. *The Structure in Person's Implicit Taxonomy of Emotions*. Journal of Research in Personality, 1982.

[41] J. Salojärvi, K. Puolamäki, and S. Kaski. *Implicit Relevance Feedback from Eye Movements*, volume Volume 3696/2005 of *Artificial Neural Networks: Biological Inspirations – ICANN 2005*. Springer Berlin, 2005.

[42] T. Saracevic. Relevance reconsidered. information science: Integration in perspectives. *Proceedings of the Second Conference on Conceptions of Library and Information Science, Copenhagen, Denmark*, pages 201–218, 1996.

[43] K. R. Scherer. *Appraisal considered as a process of multi-level sequential checking*. Appraisal processes in emotion: Theory, Methods, Research. Oxford University Press, New York and Oxford, k. r. scherer, a. schorr, & t. johnstone (eds.) edition, 2001.

[44] K. R. Scherer. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729, December 2005.

[45] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2007. ACM Press.

[46] A. F. Smeaton and S. Rothwell. Biometric responses to music-rich segments in films: The cdvplex. In *Seventh International Workshop on Content-Based Multimedia Indexing*, pages 162–168, 2009.

[47] M. Smeulders, A., M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 22, pages 1349–1380. IEEE Computer Society, 2000.

[48] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun. Affective ranking of movie scenes using physiological signals and content analysis. In *MS '08: Proceeding of the 2nd ACM workshop on Multimedia semantics*, pages 32–39, New York, NY, USA, 2008. ACM.

[49] C. Tenopir, P. Wang, Y. Zhang, B. Simmons, and R. Pollard. Academic users' interactions with sciencedirect in search tasks: Affective and cognitive behaviors. *Information Processing and Management: an International Journal*, 44(1):105–121, 2008.

[50] J. Urban and J. Jose, M. Evaluating a workspace's usefulness for image retrieval. *Journal of Multimedia Systems*, 12(4-5):355–373, 2007.

[51] R. Valenti, N. Sebe, and T. Gevers. Facial expression recognition: A fully integrated approach. *Image Analysis and Processing Workshops, 2007. ICIAPW 2007. 14th International Conference on*, pages 125–130, Sept. 2007.

[52] A. Vinciarelli, N. Suditu, and M. Pantic. Implicit human-centred tagging. In *Proceedings of IEEE International Conference on Multimedia and Expo*, 2009.

[53] S. Vrochidis, P. King, L. Makris, A. Moumtzidou, V. Mezaris, and I. Kompatsiaris. Mklab interactive video retrieval system. In *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 563–564. ACM, 2008.

[54] P. Wang, B. Hawk, W., and C. Tenopir. Users' interaction with world wide web resources: an exploratory study using a holistic approach. *Information Processing and Management: an International Journal*, 36(2):229–251, 2000.

[55] G. M. Wilson and M. A. Sasse. Listen to your heart rate: counting the cost of media quality. pages 9–20, 2000.