Affective Feedback: An Investigation into the Role of Emotions in the Information Seeking Process

Ioannis Arapakis
Department of Computing Science
University of Glasgow
Lilybank Gardens
Glasgow, G12 8QQ
arapakis@dcs.gla.ac.uk

Joemon M. Jose
Department of Computing Science,
University of Glasgow
Lilybank Gardens
Glasgow, G12 8QQ
jj@dcs.gla.ac.uk

Philip D. Gray
Department of Computing Science,
University of Glasgow
Lilybank Gardens
Glasgow, G12 8QQ
pdg@dcs.gla.ac.uk

ABSTRACT

User feedback is considered to be a critical element in the information seeking process, especially in relation to relevance assessment. Current feedback techniques determine content relevance with respect to the cognitive and situational levels of interaction that occurs between the user and the retrieval system. However, apart from real-life problems and information objects, users interact with intentions, motivations and feelings, which can be seen as critical aspects of cognition and decision-making. The study presented in this paper serves as a starting point to the exploration of the role of emotions in the information seeking process. Results show that the latter not only interweave with different physiological, psychological and cognitive processes, but also form distinctive patterns, according to specific task, and according to specific user.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process; H.5 [**Information Interfaces and Presentation**]: User Interfaces

General Terms

Experimentation, Human Factors, Measurement.

Keywords

Relevance feedback, facial expression analysis, affective interaction.

1. INTRODUCTION

User feedback is considered to be a critical element in the information seeking process [7]. A key feature of the feedback cycle is relevance assessment that has progressively become a popular practice in web searching activities and interactive information retrieval (IR). The value of relevance assessment lies in the disambiguation of the user's information need, which is achieved by applying various feedback techniques. Such techniques vary from explicit to implicit and help determine the relevance of the retrieved documents.

The former type of feedback is usually obtained through the explicit and intended indication of documents as relevant (positive feedback) or irrelevant (negative feedback). Explicit feedback is a robust method for improving a system's overall retrieval performance and providing better query reformulations [14] at the expense of users' cognitive resources [2]. A number of studies [14][15] provide evidence that explicit relevance feedback based techniques (e.g. term suggestion) are generally desirable, even though they are rarely being applied during an information seeking process [2]. Furthermore, explicit feedback techniques suffer from a significant trade-off, between the users perusing documents because the system expects them to do so and because they actually exhibit a genuine interest towards their content. Eventually, as the task complexity increases the cognitive resources of the users stretch even thinner, turning the process of relevance assessment into a non-trivial task [2].

On the other hand, techniques that fall under the category of implicit feedback tend to collect information on search behavior in a more intelligent and unobtrusive manner. By doing so, they disengage the users from the cognitive burden of document rating and relevance judgments. Information-seeking activities such as reading time, saving, printing, selecting and referencing [20][15][33] have been all treated as indicators of relevance, despite the lack of sufficient evidence to support their effectiveness [25]. From the findings provided by Kelly and others [10][11][12], it is evident that several reliability issues arise when attempting to infer relevance feedback based on observable search behaviors, simply because what can be observed does not necessarily correspond to the underlying intention. Even though implicit feedback measures are considered attractive and useful alternatives, especially when large amounts of data can be obtained very easily, they are not always inherently so. According to Kelly and Belkin [11], implicit feedback measures that use interaction with the full content of documents can often be unreliable, difficult to measure and interpret.

As shown, both categories of feedback techniques determine document relevance with respect to the cognitive and situational levels of the interactive dialogue that occurs between the user and the retrieval system [29]. However, this approach does not account for the dynamic interplay and adaptation that takes place between the different dialogue levels, but most importantly it does not consider the affective dimension of interaction. Users interact with intentions, motivations and feelings besides real-life problems and information objects, which are all critical aspects of cognition and decision-making, as shown by recent studies

[5][28][27]. Therefore, there is a need to reconsider relevance feedback with respect to what occurs on the affective level of interaction as well.

In an earlier study, Kuhlthau [16] proposed a six-stage model for the information search process (ISP), based on observations of the search behaviour of high school students. Kuhlthau's findings indicate that the information search process is an integration of three dimension of the human experience, namely: (i) affective, (ii) cognitive, and (iii) physical. Most importantly, her work brought attention to the fact that feelings such as uncertainty, confusion, anxiety and other, play an important role in the search process, and that their presence should be considered as natural and necessary. Further evidence that support the interrelation of affective, cognitive and physical behaviors was delivered by Nahl and Tenopir [24], and Nahl [21][22]. More in specific, Nahl [21] found that the affective component of information search behavior can regulate cognitive processing through a hierarchical organization of goals, which is prescribed by both individual and cultural elements. A number of other studies [34][17][36][13][23][3] also examined the affective aspects of search behaviour. The reported evidence indicates that the cause of certain emotions can relate to system, search strategy and search results [34], as well as content design and aesthetics [17]. The same studies have also shown the influence of affect on user motivation [22], performance [36][22][13][23] and satisfaction [3].

Nevertheless, very limited research has been done in relation to the role of retrieved content as emotional stimuli and its impact on user affective behaviour. In [18], Lopatovska and Mokros performed a study where users had to evaluate a number of websites with respect to a given search task. These evaluations were expressed in the form of two measures of affective value, namely: Willingness-to-Pay (WTP) and Experienced Utility (EU).The results of the study indicate that both WTP and EU reflect user's rational and emotional perception. The former is related to the website's perceived usefulness in solving the task at hand, while the latter to the general interest in its content and its aesthetic features. In [19], Mooney et al. 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 biometric measurements (GSR, skin temperature, etc.). The study provides some initial evidence that support the use of biometrics in the latter context.

The current work, following the example set by [19], investigates the role of emotions in the information seeking process and the potential impact of task difficulty on users' emotional behaviour. Most importantly, it introduces a new approach to the detection and quantification of affective information, which can be potentially applied in future studies to analyze search behaviour at relevance assessment level. However, due to the exploratory character of our study a full analysis of the collected data is beyond the scope of this paper and remains to be published in subsequent work.

2. EXPERIMENTAL METHODOLOGY

Even though physiological response patterns and affective behavior are observable, there are no objective methods of measuring the subjective experience [32]. Most researchers simply ask the participants to provide a description of their emotional experience using a combination of think-aloud protocols [34][24] and forced-choice [13][36] or free-response reports. In other cases [8][19] affective behavior is decomposed and examined through the application of a multi-modal analysis of different communication channels [9]. However, those studies suffer from a significant trade-off between the participants being aware they are recorded (open recording) and not possessing that knowledge, therefore acting spontaneously (hidden recording).

By definition an experimental study introduces the participants to an artificial situation that takes place at a laboratory setting, therefore lacking the ecological validity of a naturalistic study. In addition, when recording facial expressions several critical issues arise [30]. Firstly, emotional expressions are highly idiosyncratic in nature and may vary significantly from one individual to another (depending on personal, familial or cultural traits). Secondly, spontaneous expressive behavior may not be easily elicited, especially when participants are aware of being recorded. Finally, while interacting with researchers and other authorities the participants may intentionally try to mask or control their emotional expressions, in an attempt to act in appropriate ways. While taking into consideration the above factors we devised a user study that mitigated most of the unwanted effects. In our approach we: (i) employed a facial expression recognition system of reasonably robust performance and accuracy across all individuals, (ii) applied hidden recording, thus increasing the chance of observing spontaneous behaviour, and (iii) made our presence in the laboratory setting as unobtrusive as possible. Our primary goal was to create sufficient ground truth where facial expressions would correspond to the current emotional state of every participant.

2.1 Design

This study used a repeated-measures design. There was one independent variable: task difficulty (with three levels: "T_1: easy", "T_2: very difficult" and "T_3: practically impossible"). The levels were controlled by assigning topics with the appropriate number of relevant documents within the corpus (more than 100, less than 20, one or zero), therefore improving or decreasing the chance of one finding relevant documents accordingly. The dependent variables were divided into three subgroups, namely: (i) task, (ii) search process, and (iii) emotional experience. Among the many aspects of each subgroup, we measured perceived task difficulty, task complexity, search information need vagueness, and other.

2.2 Participants

Twenty-four participants of mixed ethnicity and educational background (9 Ph.D. students, 3 MSc students and 12 BSc students) applied for the study through a campus-wide ad. The participants were from 11 different programs: bioinformatics, biology, business administration, computing science, electrical engineering, geology, international studies, international communication, law, mathematical science and sociolinguistics. They were all proficient with the English language (9 native, 12 advanced and 3 intermediate speakers). Of the 24, 12 were male and 12 were female. All participants were between the ages of 18 and 45, and free from any obvious physical or sensory impairment. They had a mean of 8.25 years of searching experience and 23 out of 24 claimed to have been using at least one popular (among many) search service in the past.

2.3 Apparatus

For our experiment we used a desktop computer, equipped with a conventional keyboard and mouse. A *Live! Cam Optia AF* web camera with a 2.0 megapixels sensor was also mounted on top of the computer screen and was used to film the participants' expressions. To conceal the operation of the camera we made it look as inactive by exposing a disconnected power cable that apparently belonged to it.

2.3.1 Logging Software

The desktop computer was equipped with *BB FlashBack* (http://www.bbsoftware.co.uk) screen recorder that unobtrusively monitored and recorded participants' desktop activity. Information such as URLs visited, start, finish and elapsed times for interactions, keystrokes and clicks were also recorded and stored in a data file located on the desktop computer. For the capturing of participants' facial expressions we used the default recording software that was provided with the web camera. The video recordings were executed in stealth mode, for the duration of each search task, and captured all possible facial expressions. The collected data were then used to determine the probability of each expression (per key-frame) matching any of the detectable, by the facial expression analyzer, emotions and store the scores into a log file. The video recordings were also retained for further analysis in combination with the screen recordings (picture in picture effect), to infer conclusions about the source of emotional stimuli (recognition of a relevant document, a search query that produced no interesting results, etc).

2.3.2 Questionnaires

The participants completed an Entry Questionnaire at the beginning of the study, which gathered background and demographic information, as well as previous computer and searching experience. The information obtained from the Entry Questionnaire was used to characterize subjects, but not in subsequent analysis. Post-search Questionnaires were also administered at the end of each task, to elicit participants' viewpoint on certain aspects of the search process. The questions were divided into three sections that covered the search process, the encountered task and participants' emotional experiences. The last section, which enquired information regarding the experienced emotional episodes, was an adaptation of the Geneva Appraisal Questionnaire (GAQ) [31]. GAQ has been developed by the members of the Geneva Emotion Research Group, on the basis of Klaus R. Scherer's Component Process Model of Emotion (CPM). It consists of 35 questions, which have been divided into eight categories, namely: (i) occurrence of the emotional experience, (ii) general evaluation of the event, (iii) characteristics of the event, (iv) causation of the event, (v) consequences of the event, (vi) reactions with respect to the real or expected consequences, (vii) intensity and duration of the emotional experience, and (viii) verbal description of the emotional experience. Its purpose is to assess, as much as possible, through recall and verbal report the results of a participant's appraisal process in the case of an emotional episode. 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 produced the emotional episode, as well as details about what happened and the consequences it had for the participant. Out of the 35 questions of GAQ we used only 18 (4-9, 18-23, 25, 29 and 31-34) and retained the structure of categories ii, iii, v, vii and viii in our Post-search Questionnaire. In general, by decomposing the search process to a set of parameters and addressing them through different questions, we were able to identify how the different levels of our independent variable influenced them. Finally, the participants completed an Exit Questionnaire at the end of the study that gathered information on the perceived task and information need ambiguity, as well as their views of the importance of affective feedback, with respect to usability and ethical issues.

2.3.3 Search User Interface

For the completion of the search tasks we used Indri, which is an open source search engine from the Lemur project[1]. Indri is a flexible and reliable tool that provides its own complete structural query language, as well as a search interface. The query environment interface was modified to appear as one of the popular search interfaces, under the name *Chest of Knowledge*. This modification was made purposely to exploit participants' familiarity with existing search services. One of the main reasons for choosing the Indri search engine was its ability to parse TREC newswire and web collections and return results in the TREC standard format. The main disadvantage that we encountered was the complexity of the query language structure.

2.3.4 Test Collection & Search Tasks

For the indexing we used *TREC 9 (2000) Web Track*, which is a 1.69 million document subset of the VLC2 collection, of 10 gigabyte size. WT10g has been improved by eliminating many of the binary and non-English pages normally found in web crawls [1]. According to Borlund [4], TREC topics and simulated information need situations share a similar structure, which consists of a number of sections. However, in terms of limiting the area of searching a TREC topic appears to be more useful than a simulated information need situation. The basic assumption behind the topic frame is that an information need is considered as static and well defined, which provides an objective measure of recall. The simulated information need situation, however, does not introduce such artificial limitations. The only element that is considered static is the simulated task situation, i.e. the known reason for the indicative request. This allows for personal interpretations of the information need, which can lead to modifications of their initial or later search queries.

In this study, even though we retained the original content of the TREC topics, we presented them using the structural framework of the simulated information need situations. By doing so, we introduced short *cover stories* that helped us describe to our participants the source of their information need, the environment of the situation and the problem to be solved, thus facilitating a better understanding of the search objective [4]. In addition, we introduced a layer of realism to the search tasks, while preserving well-defined relevance criteria (as the latter are specified by each TREC topic description).Based on our criterion for defining task difficulty, we formulated two different scenarios for each level and allowed our participants to complete the one they considered more interesting.

[1] http://www.lemurproject.org/

2.3.5 Facial Analysis Software

The video recordings were edited using *Adobe Premiere Pro CS3*. The beginning and ending sections of each recording were trimmed off, in order to isolate the parts of the videos that showed the participants working actively on their search tasks. Those parts were afterwards synchronized with the screen recordings and a *picture-in-picture* effect was applied, followed by manual annotation of each session. The facial expression analysis was performed using eMotion [30], a facial expression recognition system developed by Roberto et al. [35]. eMotion follows a model-based approach, where an explicit 3-dimentional wireframe model of the face is constructed. Once certain facial landmark features are detected (such as the eyebrows, the corners of the mouth, etc.), a face model consisting of a number of surface patches is warped to fit them. Upon the construction of the model, head motion or facial deformations can then be tracked and measured in terms of motion-units (MU's).

The version of the facial expression recognition system that we used applies a generic classifier that has been developed from a subset of the Cohn-Kanade database. Its main advantage is that it performs reasonably well across all individuals, independently of ethnicity-specific features. We restrain from claiming that such characteristics are not of importance, especially on an interpretation level. Nevertheless, even though the classifier won't give the optimal performance it is still reasonably robust to most of the variation introduced from mixed-ethnicity groups, since it has learned statistically by studying many different individuals. Another important issue is emotion extraction. eMotion applies a static classification scheme, which entails the processing of each frame independently from its neighboring frames and classifies it to one of the facial expression categories. Static classification is considered more error-prone and unreliable [6]. However, it does not require an extensive knowledge of the object of analysis and is generally faster and simpler to implement.

Finally, facial expression recognition systems do not take into consideration context and, therefore, cannot perform a context-dependent interpretation of the data [9]. Fasel and Luettin [6] argue that facial expression recognition should not be confused with human emotion recognition. Even though the former deals with the classification of facial motion into distinct emotion categories, human emotions are the results of various intrinsic or extrinsic factors and their state may or may not be revealed through a number of channels. This argument, however, does not negate the fact that judgments based on facial expressions and other behavioral cues are far more accurate than those that are based on the body or the tone of the voice alone [26].This in turn suggests that affective information conveyed by the visual channel can be crucial to human judgment and offer valuable insights about the emotional state of the observed person. Unfortunately, the same kind of information cannot be inferred from questionnaires, since people tend to be less spontaneous and expressive. To conclude, the results from the automatic facial expression analysis have been used only as cues for emotion recognition and not as the ground truth itself.

While conducting this study we also took into consideration several other issues, such as occlusion, illumination conditions, and other, which could have introduced noise to the analysis. We are aware that there is no such thing as a flawless data set and we refrain from claiming that our data are completely accurate. Nevertheless, we believe that we have accumulated a reasonable amount of evidence to support our arguments. For a more detailed presentation of the above issues the reader is referred to [9][26].

2.4 Procedure

The user study was carried out in the following manner. The formal meeting with the participants occurred in the laboratory of the researcher. At the beginning of each session the participants were informed about the conditions of the experiment, both verbally and through a Consent Form, and then completed an Entry Questionnaire. The session proceeded with a brief training on the use of the search interface. Also, to ensure that the participant's face would be visible on the web camera we encouraged them to keep a proper posture, while interacting with the search interface, by indicating health and safety measures. Every participant completed three search tasks in total. In each search task they were handed two scenarios, both of the same level of difficulty, and were asked to proceed with the one they preferred the most. Each scenario description provided well-defined criteria for document relevancy. To negate the order effects we counterbalanced the task distribution by using a Latin Squares design. The participants were asked every time to bookmark as many relevant documents as possible (with a minimum number of 10 relevant documents) and were given 10 minutes to complete the scenario of their choice, during which they were left unattended to work. At the end of each task the participants were asked to complete a Post-search Questionnaire. An Exit Questionnaire was also administered at the end of each session along with a second Consent Form, which provided a detailed explanation of the unknown study conditions and was granting us permission to retain the video recordings for future analysis. The participants were encouraged to ask questions and were notified that they had the right to withdraw, without their legal rights or benefits being affected. In addition, all data gathered on them would be instantly and permanently destroyed. Finally, the participants were asked to sign a Payment Form, prior to receiving the participation fee of £10.

3. RESULTS

This section presents the experimental results of our study, based on 72 searches carried out by 24 participants. We collected questionnaire data on three aspects of the information seeking process, namely: (i) tasks, (ii) search process, and (iii) emotional experience. A 5-point Likert scale was used in all questionnaires, where high scores represent a stronger perception and low scores represent a weaker perception in our analysis. Friedman's ANOVA and Pearson's Chi-Square test were used to establish the statistical significance ($p < .05$) of the differences observed among the three tasks (T_1, T_2, and T_3). When a difference was found to be significant the Wilcoxon Signed-Ranked Test was applied to isolate the significant pair(s), through multiple pair-wise comparisons. To take an appropriate control of Type I errors we applied a Bonferroni correction, and so all effects are reported at a .016 level of significance. Additionally, we gathered performance data based on a preliminary analysis of the video recordings (facial expressions and screen recordings). One-Way Repeated Measures ANOVA was used to verify any statistically significant differences ($p < .05$) in participants' search performance. When a difference was found to be significant the Bonferroni post hoc test was applied to isolate the significant pair(s).

3.1 Tasks

Table 1 shows the means and standard deviations for participants' assessment of the task difficulty. It appears that there is a trend on the perceived level of difficulty among tasks T_1 to T_3, with T_1 considered as the easiest. Friedman's ANOVA was applied to evaluate the effect that the manipulation of the actual task difficulty had on the perceived task difficulty. The results indicate that participants' perception of task difficulty was significantly affected ($\chi^2(3, N{=}24) = 21.900, p < .05$). The post hoc tests show that the differences between T_1 & T_2 ($Z = -3.934, p < .016$) and T_1 & T_3 ($Z = -3.419, p < .016$) are statistically significant, but the same condition does not apply for T_2 & $T_{3.}$ A two-way repeated measures ANOVA analysis was also conducted on the performance data. This revealed that the number of bookmarked documents was affected by the level of task difficulty, $F(1.43, 31.50) = 51.7, p < .05, r = .70$. Mauchly's test indicated that the assumption of sphericity had been violated, $\chi^2(2) = 10.61, p < .05$, therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .71$). Bonferroni post hoc tests revealed a significant difference in the number of bookmarked documents between T_1 & T_2 and T_1 & T_3 ($p < .016$). No other comparison was found significant. Table 1 also shows participant's subjective assessment on the complexity and ambiguity of the three tasks. Friedman's ANOVA test reveals a significant difference for task complexity, but does not indicate the same for task ambiguity. The Wilcoxon tests show that the difference in complexity is significant for pairs T_1 & T_2 ($Z = -3.333, p < .016$) and T_1 & T_3 ($Z = -2.753, p < .016$).

Table 1. Descriptive statistics on task aspects

Task	Difficulty		Complexity		Ambiguity	
	M	SD	M	SD	M	SD
T_1	1.5417	0.8330	1.5000	0.5898	1.3333	0.6370
T_2	3.3333	1.0495	2.5417	0.9771	1.6667	0.9168
T_3	3.1667	1.4346	2.4583	1.3181	1.3333	0.4815

Table 2. Descriptive statistics on task aspects

Task	Information need clarity		Easiness of query formulation	
	M	SD	M	SD
T_1	4.0417	0.9990	4.0000	1.1547
T_2	3.5417	0.9315	2.8636	0.9902
T_3	4.3750	1.0135	3.0455	1.1329

The means and standard deviations for participants' understanding of their simulated information need, as well as the perceived easiness of formulating appropriate query statements, are presented in Table 2. The participants were asked to provide their assessments through the following questions: (i) "How well defined was your information need for the current task?", (ii) "It was easy to formulate queries for this topic (Range: 1-5, Lower = Disagree)". Friedman's ANOVA shows that the difference among the three tasks is significant for both information need clarity ($\chi^2(3, N{=}24) = 10.314, p < .05$) and easiness of query formulation ($\chi^2(3, N{=}24) = 14.514, p < .05$). For the former variable, the post hoc test did not indicate a significant difference among any of the tasks, while for the latter variable it revealed a significant pair-wise difference for T_1 & T_2 ($Z = -3.337, p < .016$) and T_1 & T_3 ($Z = -2.915, p < .016$) only.

3.2 Search Process

Similarly to the task, we examined the effect of our independent variable to the search process. Table 3 shows the means and standard deviations of participants' subjective assessment on search process difficulty, interest and fatigue. The Friedman's ANOVA test shows that search difficulty differs significantly across all tasks ($\chi^2(3,N{=}24) = 26.690, p < .05$). However, the post-hoc tests show that only the pairs T_1 & T_2 ($Z = -3.778, p < .016$) and T_1 & T_3 ($Z = -4.028, p < .016$) have a significant difference. Search interest was also found by Friedman's ANOVA to have a significant difference ($\chi^2(3, N{=}24) = 9.896, p < .05$). The Wilcoxon tests reveal that only the difference between T_1 & T_3 ($Z = -2.973, p < .016$) is statistically significant. Finally, the levels of perceived fatigue across the three tasks appear to differ significantly ($\chi^2(3, N{=}23) = 14.986, p < .05$). The Wilcoxon tests indicate a significant difference for pair-wise comparisons of tasks T_1 & T_2 ($Z = -2.430, p < .016$) and T_1 & T_3 ($Z = -3.451, p < .016$).

Table 3. Descriptive statistics on search process aspects

Task	Difficulty		Interest		Fatigue	
	M	SD	M	SD	M	SD
T_1	1.8333	0.9630	3.8333	1.2038	2.5217	0.7304
T_2	3.6667	1.2394	3.3750	1.2445	3.2174	0.9023
T_3	3.8750	0.8998	2.8333	1.0901	3.5217	0.7902

Table 4. Descriptive statistics on emotional experience

Task	Unpleasantness of stimuli		Intensity of emotion		Effort to mask emotional expression	
	M	SD	M	SD	M	SD
T_1	1.8750	1.0759	3.2500	1.2597	2.3636	1.1358
T_2	2.9583	1.1220	2.9167	0.8297	2.1364	0.8335
T_3	3.1250	1.0759	3.2500	1.1515	2.0455	0.8438

3.3 Emotional Experience

To evaluate the progression of the emotional patterns across the three tasks we asked the participants to self-assess the emotional episodes they experienced during the study. Table 4 shows a summary of some of the most important aspects of the emotional episodes, such as the perceived unpleasantness of the stimuli, the intensity of the experienced emotion, as well as the amount of effort that the participants put to control or mask their emotional expressions. Friedman's ANOVA test was used on all three variables. We found a significant difference only for the unpleasantness of the stimuli, across the different conditions ($\chi^2(3, N{=}24) = 14.364, p < .05$). The Wilcoxon Signed-Ranked tests that followed up this finding also reveal that the significant difference lies between pairs T_1 & T_2 ($Z = -2.932, p < .016$) and T_1 & T_3 ($Z = -3.552, p < .016$). A major goal of this study was to confirm the occurrence of emotions during an information seeking process. The pie charts in Figure 1 illustrate the pattern of distribution of the most intense emotions, as the latter were reported for each task by the 24 participants. The first pie chart reveals that *happiness* and *irritation* were the most intense emotions, among all other reported emotions in task one (T_1), followed by *sadness*, *pleasure* and *surprise* respectively. The second pie chart shows a different distribution, with *irritation* being reported by half of the participants as the most intense emotion during the second task (T_2). Other emotions such as

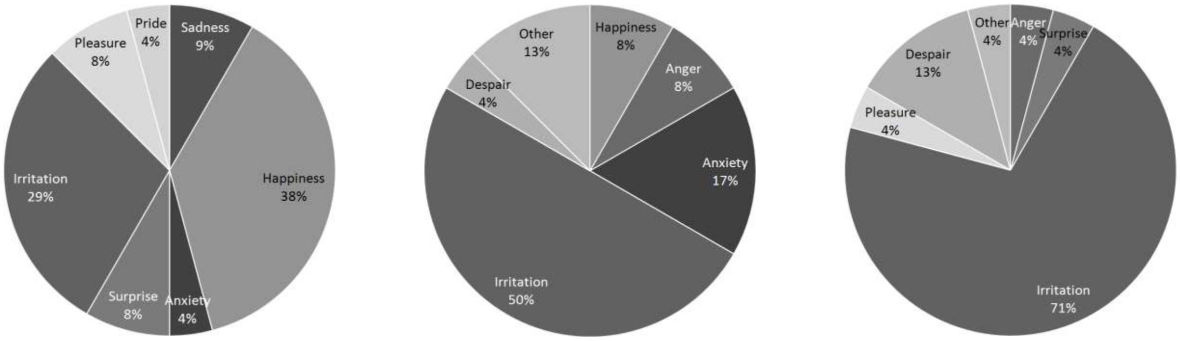

Figure 1. Distribution of emotions for Tasks 1-3 (order of appearance: left to right)

anxiety, *anger* and *happiness* were also reported on at a lesser rate. Finally, the third pie chart indicates that *irritation* was the dominant *emotion* for the third task (T_3), accompanied by other emotions, such as *despair*, *anger*, *surprise* and *pleasure*. Pearson's Chi-Square test was also applied and revealed a significant variation in the distribution of *irritation* ($\chi^2(2, N=24) = 8.33, p < .05$) and *happiness* ($\chi^2(2, N=24) =11.4, p < .05$), across the three tasks.

3.4 Automatic Facial Expression Analysis

In this section we present the preliminary results of the facial expression analysis that we performed using eMotion [30]. Each session was processed separately and the data were stored in a log file, which was labeled with the participant's unique ID, task number and order number. The log data display for each key-frame of the video recording the probability of the detected facial expression (assuming there was one) corresponding to any of the seven basic emotional categories that eMotion can recognize (a higher percentage score corresponds to greater confidence in the classification of the detected facial expression and to higher intensity). For each log file we counted the number of key-frames, per emotion, that received a probability greater than .90. We deliberately set a high threshold to exclude emotions that were detected with low probability scores; therefore minimize the noise in our data. We then divided these scores with the total number of key-frames of the video sequence, to normalize its contribution to the average values across all videos and per task. The bar chart in Figure 2 shows the average values of the aggregates across all participants, for tasks T_1, T_2 and T_3. Additionally, we chose a random participant and examined the log file data to acquire a micro view of the emotional variation, across tasks T_1, T_2 and T_3. The bar chart in Figure 3 illustrates the frequency of the seven basic emotions, as the latter were detected by eMotion for the selected sample (the scores were again filtered using a .90 threshold). The dissimilar distribution of scores makes evident the emotional blend that characterizes the different level of difficulty, under which each task was conducted. Elements of interest are the type of emotion, as well as its frequency of occurrence.

4. DISCUSSION

Overall, the manipulation of the task difficulty, through the exercise of control on the availability of relevant documents, appeared to have a significant effect on several aspects of the information seeking process. Statistically significant differences were found between perceived task difficulty and complexity, as well as information need clarity and easiness of query formulation. In all cases, the post-hoc tests indicated a significant difference for pairs T_1 & T_2 and T_1 & T_3. This finding suggests that, from the viewpoint of the participants, the degree of variation of the above measures did not prove consistent across all pair-wise task comparisons (specifically for tasks T_2 and T_3) and, therefore, was not easily perceptible in some of the cases.

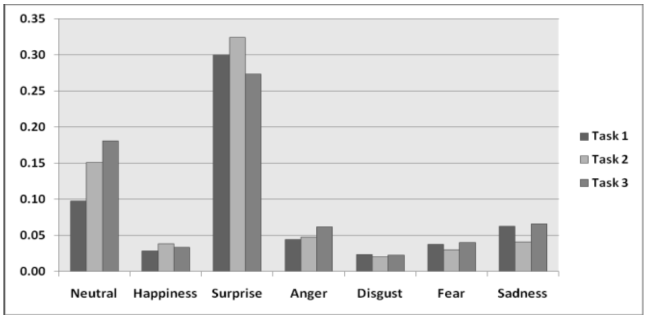

Figure 2. Average scores of detected emotions, across all users, for Tasks 1-3

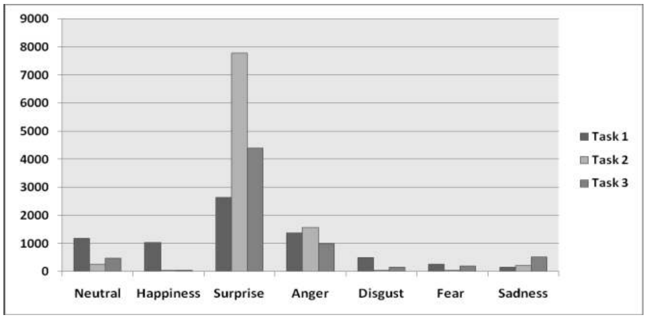

Figure 3. Aggregated scores of detected emotions from a random sample, for Tasks 1-3

One could argue that the number of relevant documents in the collection is not a very reliable measure for the task difficulty of a search topic, since retrieval is not a random process but rather a factor of many things (including query statements, indexing language and the retrieval mechanism). As a result, even when dealing with an easy task, the retrieval system can still collect poor results if given a query statement of poor quality. However, the results from the performance data analysis, presented in section 3.1, confirm that there were indeed perceived differences in the topic difficulty. The post hoc tests that followed up this finding revealed a significant difference in the number of bookmarked documents between T_1 & T_2 and T_1 & T_3. No statistically significant effect on task ambiguity was found for any participants, suggesting that the task descriptions were clearly defined. An examination of the mean scores of Table 3 clearly distinguishes some tasks from others, most notably that reported search difficulty and complexity escalated as the task difficulty increased. Again, pair-wise differences were found significant for tasks T_1 & T_2 and T_1 & T_3. These differences indicate an analogy between the above factors and the difficulty of the task at hand, most likely due to the mutual interaction between task and search process. A similar finding applies for search interest, which increased in an inverted manner in comparison with task difficulty. A statistically significant difference was evidenced in the post-hoc tests only for pair T_1 & T_3. This suggests that easy tasks promoted a more engaging and stimulating experience, contrary to difficult tasks that had a negative effect on participants' level of interest. The analysis also shows that the participants put very little effort to mask their emotional expressions and, therefore, we can reasonably assume that these were spontaneous and genuine. This behaviour was consistent across all three tasks. The intensity of the experienced emotions did not vary significantly. However, the unpleasantness of the stimuli was found significantly different between pairs T_1 & T_2 and T_1 & $T_{3,}$ revealing a trend towards negative emotional stimuli as the task difficulty arises.

Furthermore, from the pie charts in Figure 1 it is evident that some interesting patterns of emotional variation emerge. The most critical conclusion at this point is that task difficulty and complexity have a significant effect on the distribution of emotions across the three tasks. As the former increase, so do the negative emotions intensify and progressively overcast the positive ones. We hypothesize that this progression is the result of an underlying analogy between the aforementioned search factors and emotional valence, and, furthermore, that it is indicative of the role of affective information as a feedback measure, on a cognitive, affective and interactional level. Additional insights can be drawn by examining the behavior of the seven basic emotional categories, in terms of frequency, as they are illustrated in Figure 2. The average scores across the three tasks show that the least frequent occurrences were logged for *happiness*, *anger*, *disgust*, *fear* and *sadness*, with *surprise* being the most frequently expressed emotion (according to the facial expression analysis produced by eMotion). No other significant variation in the aggregated frequencies is evident between the tasks (this insight does not necessarily apply for the distribution of emotions throughout the search process, which remains to be studied). We speculate that the low frequency scores of some emotions might make them better feedback indicators, compared to other categories that exhibit higher scores. We refrain from claiming that frequently occurring emotions do not convey potentially important affective information. However, it is perhaps the rarity of the emotional stimuli that might be correlated with significant events or breakdowns throughout the search process, which makes the former group of emotions the foci of our follow-up analysis.

Finally, the bar chart in Figure 3 provides a closer peek in the aggregated frequency scores of the seven basic emotions for a single participant and the way these blend and interweave to form distinct patterns in each task. Since this is only a random sample, taken from our somewhat larger subset, we will restrain from generalizing to the whole population. Nevertheless, it constitutes a fine example of the not so apparent emotional diversity that we often fail to notice in ourselves and others.

5. CONCLUSIONS

We conducted an exploratory user study involving 24 participants and collected a set of multimodal interaction data. Several important conclusions can be drawn. Foremost among them is that emotions not only interweave with different physiological, psychological and cognitive processes during the search process, but also form distinctive patterns. These patterns might prove to be good predictors of document relevancy or indicate significant events and breakdowns that are correlated with changes in the users' knowledge state and information need. Moreover, our findings reveal that users' emotions progressively transit from positive to negative valence, as the degree of task difficulty increases. This suggests that the affective feedback should be treated differently as the task difficulty increases; and thus we should interpret the relevance indicators accordingly.

However, additional analysis must be performed in order to validate the clarity of this argument. We believe that the quality and comprehensiveness of our data can provide much insight into the role of emotions in the information seeking process. A post microscopic analysis of all logged sessions will allow us to associate the occurrence of emotions with significant stages and events in the search process, as well as facilitate a better understanding of their significance. Additionally, a simulation of feedback techniques will allow us to examine the role of emotions at relevance assessment level. Further testing of a wider range of modalities is also part of our future research. We are aware that the present study has several limitations. However, these are only the first steps into a new and unexplored domain and a full analysis of the data is beyond the scope of this paper. With this work we believe that we contribute to the exploration of the role of emotions in the search process. Furthermore, we introduce a new approach to the detection and quantification of affective information, in an attempt to reconsider relevance feedback on a cognitive as well as affective level.

6. ACKNOWLEDGMENTS

The research leading to this paper was supported by the European commission, under the contract FP6-033715 (MIAUCE Project). Our thanks to Roberto Valenti and Nicu Sebe for providing us with an evaluation version of eMotion, and Hideo Joho for useful discussions.

7. REFERENCES

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

[2] Belkin, N.J., Cool, C., Head, J., Jeng, J., Kelly, D., Lin, S.J., Lobash, L. Park, S.Y., Savage-Knepshield, P., and Sikora, C. Relevance feedback versus Local Context Analysis as term suggestion devices: In Proceedings of the Eighth Text Retrieval Conference TREC8.

[3] Bilal, D., and Kirby, J. 2002. Differences and similarities in information seeking: children and adults as web users. Inf. Process. Manage. 38, 5, 649-670.

[4] Borlund, P. 2000. Experimental components for the evaluation of interactive information retrieval systems. Journal of Documentation, 56(1), 71-90.

[5] Damasio, A. R., 1994. Descartes Error: Emotion, Reason, and the Human Brain. Gosset/Putnam Press, New York.

[6] Fasel, B., and Luettin, J., 2003. Automatic facial expression analysis: a survey. In Pattern Recognition. 36, 1, 259-275.

[7] Harman, D. 1992. Relevance feedback revisited. In Proceedings of the 15th Annual international ACM SIGIR Conference on Research and Development in information Retrieval. ACM, 1-10.

[8] Healey, J., Seger, J., and Picard, R. W., 1999. Quantifying Driver Stress: Developing a System for Collecting and Processing Bio-Metric Signals in Natural Situations.

[9] Jaimes, A., and Sebe, N. 2007. Multimodal human-computer interaction: A survey. In Comput. Vis. Image Underst. 108, 1-2, 116-134.

[10] Kelly, D., and Belkin, N. J. 2001. Reading time, scrolling and interaction: exploring implicit sources of user preferences for relevance feedback. In Proceedings of the 24th Annual international ACM SIGIR '01. 408-409.

[11] Kelly, D., and Belkin, N. J., 2002. A user modeling system for personalized interaction and tailored retrieval in interactive IR. In Proceedings of the American Society for Information Science.

[12] Kelly, D., and Teevan, J. 2003. Implicit feedback for inferring user preference: a bibliography. SIGIR Forum 37, 2, 18-28.

[13] Kim, K. 2008. Effects of emotion control and task on Web searching behavior. Inf. Process. Manage. 44, 1, 373-385.

[14] Koenemann, J., and Belkin, N. J. 1996. A case for interaction: a study of interactive information retrieval behavior and effectiveness. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Common Ground CHI '96. ACM, 205-212.

[15] Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., and Riedl, J. 1997. GroupLens: applying collaborative filtering to Usenet news. Commun. ACM 40, 3, 77-87.

[16] Kuhlthau, C. C. 1991. Inside the search process: information seeking from the user's perspective. Journal of American Society for Information Science, 42, 5, 361–371

[17] Lavie, T. and Tractinsky, N. 2004. Assessing dimensions of perceived visual aesthetics of web sites. Int. J. Hum.-Comput. Stud. 60, 3, 269-298.

[18] Lopatovska, I., and Mokros, H. B. 2008. Willingness to pay and experienced utility as measures of affective value of information objects: Users' accounts. Inf. Process. Manage. 44, 1, 92-104.

[19] Mooney, C., Scully, M., Jones, G. J. F., and Smeaton, A. F. 2006. Investigating Biometric Response for Information Retrieval Applications. ECIR 2006: 570-574

[20] Morita, M., and Shinoda, Y. 1994. Information filtering based on user behavior analysis and best match text retrieval. In Proceedings of the 17th Annual international ACM SIGIR-94.

[21] Nahl, D. 1998. Ethnography of novices' first use of Web search engines: affective control in cognitive processing. Internet Ref. Serv. Q. 3, 2, 51-72.

[22] Nahl, D. 2004. Measuring the affective information environment of web searchers. In Proceedings of the American Society for Information Science and Technology. 41, 1, 191-197.

[23] Nahl, D. 2005. Affective and Cognitive Information Behavior: Interaction Effects in Internet Use. In Proceedings 68th Annual Meeting of the American Society for Information Science and Technology (ASIST), 42.

[24] Nahl, D. and Tenopir, C. 1996. Affective and cognitive searching behavior of novice end-users of a full-text database. J. Am. Soc. Inf. Sci. 47, 4, 276-286.

[25] Nichols, D. M. 1997. Implicit Rating and Filtering. In Proceedings of the Fifth DELOS Workshop on Filtering and Collaborative Filtering.

[26] Pantic, M., and Rothkrantz, L. J. M., 2003. Toward an Affect-Sensitive Multimodal Human-Computer Interaction. In Proceedings of the IEEE. 91, 9, 1370--1390.

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

[28] Picard, R. W., 2001. Building HAL: Computers that Sense, Recognize, and Respond to Human Emotion. In Society of Photo-Optical Instrumentation Engineers.

[29] Saracevic, T. 1975. Relevance: A review of and a framework for the thinking on the notion in Information Science. Journal of American Society for information Science, 26, 321-343.

[30] Sebe, N., Lew, M. S., Sun, Y., Cohen, I., Gevers, T., and Huang, T. S. 2007. Authentic facial expression analysis. Image Vision Comput. 25, 12, 1856-1863.

[31] Scherer, K. R. 2001. Appraisal considered as a process of multi-level sequential checking. Appraisal processes in emotion: Theory, methods, research (pp. 92-120). New York and Oxford: Oxford University Press.

[32] Scherer, K. R., 2005. What are emotions? And how can they be measured? In Social Science Information. 44, 4, 695-729.

[33] Seo, Y., and Zhang, B. 2000. Learning user's preferences by analyzing Web-browsing behaviors. In Proceedings of the Fourth international Conference on Autonomous Agents 2000. AGENTS '00. ACM, 381-387.

[34] Tenopir, C., Wang, P., Zhang, Y., Simmons, B., and Pollard, R. 2008. Academic users' interactions with ScienceDirect in search tasks: Affective and cognitive behaviors. Inf. Process. Manage. 44, 105-121.

[35] Valenti, R., Sebe, N., and Gevers, T. 2007. Facial Expression Recognition: A Fully Integrated Approach. 14th International Conference on Image Analysis and Processing Workshops. ICIAPW 2007. 125-130.

[36] Wang, P., Hawk, W. P., and Tenopir, C. 2000. Users' interaction with World Wide Web resources: an exploratory study using a holistic approach. Information Processing & Management, 36, 2, 229-251.