A Comparison of General vs Personalised Affective Models for the Prediction of Topical Relevance*

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

ABSTRACT

Information retrieval systems face a number of challenges, originating mainly from the semantic gap problem. Implicit feedback techniques have been employed in the past to address many of these issues. Although this was a step towards the right direction, a need to personalise and tailor the search experience to the user-specific needs has become evident. In this study we examine ways of personalising affective models trained on facial expression data. Using personalised data we adapt these models to individual users and compare their performance to a general model. The main goal is to determine whether the behavioural differences of users have an impact on the models' ability to determine topical relevance and if, by personalising them, we can improve their accuracy. For modelling relevance we extract a set of features from the facial expression data and classify them using Support Vector Machines. Our initial evaluation indicates that accounting for individual differences and applying personalisation introduces, in most cases, a noticeable improvement in the models' performance.

Categories and Subject Descriptors

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

General Terms

Experimentation, Human Factors, Performance

1. INTRODUCTION

The main challenge information retrieval (IR) systems face nowadays originates from the semantic gap problem: the semantic difference between a user's query representation and the internal representation of an information item in a collection. Although 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 retrieval systems to locate potentially relevant items, produce results that do not address the users' true needs.

To deal with information need uncertainty IR systems have employed in the past a range of feedback techniques, which vary from explicit [13, 21] to implicit [1, 6]. The notion of explicit feedback was present from the early years of IR, but it soon became apparent that users could not cope with the cognitive burden of explicit relevance judgments. Alternative paths had to be discovered, which led to the unobtrusive, yet less robust, implicit feedback techniques [12, 15]. Even though this was a step towards the right direction a need to personalise and tailor the search experience to the user-specific needs was progressively made evident.

Personalisation emerged as an appealing technique in dealing with the issues caused by the variation of online behaviour and the individual differences observed in user interests, information needs, search goals, difficulties encountered, and other. To apply personalisation an IR system must initially employ a modelling technique that will capture certain user characteristics. At a later stage, information filtering is performed to refine the aggregated information and adjust the system's responses to accommodate users' needs, thus providing a more personalised experience.

Several attempts have been made in the past to develop user models, using implicit feedback. In [16], Oard and Kim define a set of application-specific observable behaviours (examination, retention, etc.) and introduce the concept of learning user interests and building user profiles from implicit data. In [20], Puolamäki et al. combine implicit feedback with explicitly created user profiles. In the latter work, the authors use mixture models to combine different sources of relevance judgments. The implicit feedback information derives from eye-movement data, used in combination with a probabilistic collaborative filtering model.

In [2] the authors make the assumption that, apart from user modelling, query-specific behaviours are also important and should be considered when attempting to predict topical relevance. Following this work, Liu et al. [14] constructed user profiles based on users' search history and developed algorithms that mapped query terms to predefined categories. The latter information was used to extract users' interests and address issues related to word ambiguity. Teevan et

*The research leading to this paper was supported by the European commission, under the contract FP6-027122 (SALERO).

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

al. showed in [23] that richer representations of the user lead effectively to more accurate relevance predictions. This improvement is achieved by combining different sources of information, such as a search history, webpages visited, documents created and viewed, etc., which is used to re-rank the results obtained by a search system.

However, although the identification of user interests is a definite step, it is important to examine how these interests evolve, interact and lose focus, from a temporal perspective. In [9], Daoud et al. consider in this context the problem of search-session boundary recognition. In the above study the users were represented by long-term interests and short-term contexts, which were both essentially ontologies of semantically linked concepts. Their approach to personalisation yielded significant improvements compared to the conventional query handling paradigm.

In this paper we examine ways of personalising affective models, trained on facial expression data gathered by many individuals. Our work is limited to that of modelling users' affective behaviour and does not involve information filtering or adaptation of content. Using personalised data we adapt these models to individuals and compare their performance to a general model. For modelling relevance we extract a set of features from the facial expression data and classify them using Support Vector Machines. Our initial evaluation indicates that accounting for individual differences and applying personalisation introduces, in most cases, a noticeable improvement in the models' performance. To our knowledge, no prior work has ever applied personalisation on the affective level interaction, in the context of online information seeking.

1.1 Research Questions

The major goal of this study is to develop personalised affective models, adapted to the individual characteristics of specific users, and compare their ability to discriminate between relevant and irrelevant items against general affective models. To achieve this goal we had to expose our participants to stimuli of varied intensity. As a result, the information that we collected covered a much wider spectrum of affective behaviour and allowed the comparability of results with previous work. We, furthermore, explore different ways of combining the personalised data with the general data, to optimise the models' performance. Overall, we examined the following research hypothesis:

***H*$_1$:** By adapting a general affective model with personalised data, to a specific user, we can improve its accuracy in predicting topical relevance.

***H*$_2$:** Merging general with personalised data is more effective personalisation method compare to training separate models and applying information fusion on a decision level.

2. EXPERIMENTAL METHODOLOGY

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 analysing facial expressions several critical issues arise [22]. 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 behaviour 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 an experimental setup, similar to the one adopted in [3], 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.

2.1 Design

This study used a repeated-measures design. There were two independent variables: task difficulty (with three levels: "easy", "average", "difficult") and personalisation technique (with two levels: "adaptation" and "weighted voting"). The task difficulty levels were controlled by re-ranking the returned results to include 8 relevant - 2 irrelevant, 5 relevant - 5 irrelevant, and 2 relevant - 8 irrelevant documents, accordingly. The set of relevant documents consisted of top-ranked results, while the set of irrelevant documents consisted of bottom-ranked results. This way we improved or decreased the chances of locating relevant items among the results. The personalisation technique was controlled by adopting a different approach (mixing general with personalised data, or using them separately to train different models). The dependent variables were: (i) task (difficulty, complexity, etc.), (ii) search process, and (iii) models' performance, in terms of accuracy.

2.2 Apparatus

For our experiment we used one desktop computer, equipped with monitor, keyboard, mouse and a web-camera. The computer provided access to a custom-made search interface, which allowed the participants to perform their search tasks. The search interface was designed to re-rank the results for each submitted query, according to the level of task difficulty, without the participants being aware of it. In addition, a custom-made script logged participants' desktop actions, such as starting, finishing and elapsed times for interactions, and click-throughs. The web-camera (Creative Live! Cam Optia AF with a 2.0 megapixels sensor) was used in combination with eMotion [24], for the application of real-time facial expression analysis. Finally, we used entry-, post- and exit- questionnaires in each session.

2.2.1 Search Tasks

We prepared a number of search tasks that covered a variety of context, from entertainment to health-related issues, in order to capture participants' interest as best as possible. All tasks were performed manually, prior to the experiment, to ensure the availability of relevant documents. The search tasks were presented using the structural framework of the simulated information need situations [8]. 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. We believe that this way we facilitated a better understanding of the search objective and, in addition, we introduced a layer

Topic 1: A task of digging cheesy gossips and scandals.
Topic 2: Formulate an opinion about existing social networking sites.
Topic 3: A task of investigating, obtaining advance knowledge, or doing research on a particular sport.
Topic 4: A task of finding information regarding contraception methods.
Topic 5: A task of investigating, obtaining new knowledge, or doing research on global warming.
Topic 6: A task of planning your Christmas holidays.

Table 1: A list of the available search tasks

of realism, while preserving well-defined relevance criteria. An indicative list of the topics is presented in Table 1.

2.2.2 Search Interface

For the completion of the search tasks we used a custom-made search environment (Zoogle) that was designed to resemble the basic layout of existing search services, while retaining a minimum of graphical elements and distractions. Zoogle works on top of Yahoo! API. For every submitted query it returned a list of ten results, stripped of their title, snippet or any other metadata. This layout was intentional to ensure that the participants would not be able to judge the topical relevance of the returned documents, prior to examining them.

Even though this approach introduced our participants into artificial search situations, which differ from real-life experiences, it was a necessary trade-off for capturing affective responses exhibited towards the viewed content. In addition, we ensured that the participants viewed an equal number of relevant and irrelevant documents. This allowed us to develop balanced sets of affective data.

Zoogle applies a layered architecture approach, similar to that adopted in [5]. The first layer of the interface is dedicated for supporting any interaction that occurs during the early stages of the search process (such as query formulation and search execution). Any output generated during this phase is presented in the second layer. From there, the participants can select and preview any of the retrieved documents. The content of an item is shown in a separate panel in the foreground, which constitutes the third layer of our system.

The main purpose of this layered architecture is 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 participants' affective responses to the source of stimuli (in our case, the perused documents). This was an important aspect of our experimental methodology, since we were interested in isolating content-particular emotions. Upon examining a document, the participants had the option to either bookmark or ignore it. The first option would classify the document as relevant, while the latter as irrelevant.

2.2.3 Questionnaires

The participants completed an Entry Questionnaire at the beginning of the study, which gathered background and demographic information, and, furthermore, inquired about previous experience with online searching. 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 three sections that covered the search session, the encountered task and the returned results.

Finally, an Exit Questionnaire was introduced at the end of the study. The questionnaire gathered information on participants' views about the importance of affective feedback, with respect to usability and ethical issues. All of the questions included in the questionnaires were forced-choice type.

2.2.4 Facial Expression Recognition

Facial expressions have been associated in the past with universally distinguished emotions, such as happiness, sadness, anger, fear, disgust, and surprise [11]. Recent findings also indicate that emotions are primarily communicated through facial expressions [17] and 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 [18] but also regulate our interactions with the surrounding environment and the organisms that inhabit it.

In this study we applied real-time facial expression analysis using eMotion, an automatic facial expression recognition system with emotion-detection capabilities. The process of recognition occurred as follows: initially, eMotion would locate certain facial landmark features (eyebrows, corners of the mouth, etc.) and construct a 3-dimensional wireframe model of the face, consisting 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.

Automatic systems are an alternative approach to facial expression analysis [19] and have exhibited performance 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 [24] support our performance-related assumptions. For additional information regarding the advantages and limitations of eMotion the reader is referred to [24, 4]

2.3 Participants

Sixteen healthy participants of mixed ethnicity and educational background (8 MSc students, 4 BSc. and 4 other) applied for the study through a campus-wide ad. They were all proficient with the English language (1 native, 14 advanced, and 1 intermediate speakers). Out of 16, 7 were male and 9 were female and were between 21-32 years of age (M=25.83, SD=2.57). They had an average of 7.33 years of online search experience and all claimed to have been using at least one search service in the past (with the most popular being "Google" and "Yahoo!"). On average, the participants reported carrying out online searches once or twice a day (M=5.33, SD=0.84). 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").

2.4 Procedure

The user study was carried out in the following manner. The formal meeting with the participants took place in the laboratory setting. At the beginning of the session the participants were given an information sheet, which explained the conditions of the experiment. They were then asked to sign a Consent Form and were 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 interface, followed by a calibration of the web-camera. The participants' were told that the web-camera was used for eye-tracking purposes, thus concealing it's true operation.To ensure that their faces would be visible to the camera at all times we encouraged them to keep a proper posture, by indicating the need to stay within the visual field of the eye-tracker.

Each participant completed three search tasks, one for each level of difficulty (see Section §2.1). In every task they were handed six topics and were asked to proceed with the one they found most interesting. For each topic the subjects were given 15 minutes, during which they had to locate as many relevant documents as possible. For every submitted query the search interface would return ten results, which they were asked to evaluate one by one. If a document was judged as relevant the participants had the option to bookmark it, or otherwise ignore it and continue with the evaluation of the remaining items. Depending on the level of task difficulty ("easy", "average", "difficult") the ratio of relevant-irrelevant documents varied accordingly (the participants were unaware of this uneven distribution of relevant/irrelevant documents). To negate the order effects we counterbalanced the task distribution by using a Latin Squares design. At the end of each task, the participants were asked to complete a Post-Search Questionnaire.

An Exit Questionnaire was administered at the end of each session. The participants were informed about the unknown conditions of the study and were asked to sign a second Consent Form, which was granting us permission to retain the accumulated facial expression data. Finally, the participants were asked to sign a Payment Form, prior to receiving the fee of £10.

3. DATA ANALYSIS

Out of the 1534 browsing instances that took place during the study, 696 correspond to relevant documents and 838 correspond to irrelevant documents. Overall, we collected 440557 feature vectors, out of which 224165 are associated to relevant documents and 216392 to irrelevant documents. Our main objective was to accumulate a sufficiently rich and balanced set of affective data that would allow us to experiment with different personalisation approaches. The analysis was performed on a frame-basis.

3.1 Features

From the output of eMotion we concluded to a subset of 12 features that have directly measured values and were used to train our models. Most of these attributes have been associated in the past with important affective and cognitive processes. Even though eMotion follows the categorical approach (i.e., interprets facial expressions in terms of emotion categories) we did not employ categorical data for the training of our models. Instead, we used the motion-units (MU's) data, which is a low-level category of features very similar to Ekman's action-units (AU's) [10]. 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.

3.2 Preprocessing

We shuffled and split the data of each participant into three subsets, two of which were used for training purposes (S_1 & S_2) and one for testing (S_3). Each time we used an equal number of documents. We also resampled datasets S_1, S_2 and S_3, based on the participant with the least number of instances. This resulted in three sets with approximately the same number of feature vectors, across all participants. By balancing the training and test sets we prevented over-fitting and, additionally, compensated for the originally uneven size of the datasets. Since eMotion did not pre-processes the data we had to scale them before applying any classification method, to avoid having attributes in greater numeric ranges dominating those in smaller numeric ranges.

4. MODELS

We explore the effect of personalisation on the affective models' performance. The modelling goal is to develop affective models that can predict with reasonable accuracy the topical relevance of viewed documents. We employ sensory data that derive from facial expressions as the only implicit feedback information. From the latter signals we extract a set of features, and perform discriminant analysis, using Support Vector Machines (SVM). Additional classification techniques were evaluated in [4], but proved to be less efficient. Therefore they were omitted from this study. We do not assume anything about the relationship between these features, which we consider indicative of users' affective behaviour and topical relevance. We, rather, follow a straightforward classification approach, using the ground truth that is associated with our training data.

4.1 Support Vector Machines

We used libSVM[1], an implementation of SVM, to discriminate between two classes of documents: (i) relevant, and (ii) irrelevant. 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, based on previous work [4] that evaluated all basic SVM kernels (linear, polynomial, radial basis function, sigmoid), proved to be the optimal 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 our 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

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

Task	Easy - Difficult		Clear - Unclear		Simple - Complex		Interesting - Boring	
	M	SD	M	SD	M	SD	M	SD
T_1	1.750	0.8563	1.4375	0.6292	2.1875	1.1087	2.0000	1.1547
T_2	1.6875	0.8732	1.3750	0.6191	1.7500	1.0646	2.1333	1.1255
T_3	2.6875	1.1383	1.5000	1.0954	2.0000	1.4606	1.9375	1.2366

Table 2: Descriptive statistics on tasks

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.

4.2 Personalisation

We followed two different training approaches: (i) we merged general data, gathered from many individuals, with personalised data from a single participant and trained a single SVM model, and (ii) we used general and personalised data separately, to train two different models and combined their predictions using weighted voting.

In the first approach we used a total of 19157 instances of general data, acquired from [3], in combination with 4300 instances (in three sets of 1430 instances) per participant. For every participant we originally tested the performance of the SVM model (general model) trained on the 19157 instances against S_3 (the predestined test set of the participant). Then we retrained the model using the same general data merged with additional N instances of general data, or N^* instances of personalised data (where N or N^* equals 1430 feature vectors), and tested its performance against S_3. Finally, we repeated the same process using N+N instances of general data, or N^*+N^* instances of personalised data. This way we were able to examine if by adding personal data we improved the performance of our model more than by adding general data.

In the second approach we examined whether predictions from two different sources (general model and personalised model) could be fused, on a decision level, to determine the topical relevance of a document. For each participant we trained a personalised SVM model using the subsets S_1 and S_2. A general SVM model was also trained using the same general data as in the previous method. We then used each participant's test set (S_3) to acquire the predictions from both classifiers and combine their output using the following formula (each time with a different weighting scheme):

$$p^i_{gen} \cdot w_{gen} + p^i_{pers} \cdot (1 - w_{gen}) = p_i \quad (1)$$

Assume p^i_{gen} is the probability estimate of instance i being relevant, as given by the general model, while p^i_{pers} denotes the probability of the same instance being relevant, as determined by the personalised model. We then calculate the probability p_i of the instance i being relevant using Formula 1. Where $w_{gen} \in [0, 1]$, is the weight we assign for the prediction of the general model. The prediction p_i will then be transformed to a binary decision classifying instance i as either relevant or irrelevant, based on a predefined threshold value t. The probability estimates of both models were tested for different combinations of weights and threshold, using a step of 0.1.

5. RESULTS

In this section we present the experimental findings of our study, based on 48 search sessions that were carried out by 16 subjects. Out of the many results, we are reporting those that refer to our models and present only the questionnaire data that refer to the tasks, due to limited space. We measured the performance of all models using the standard metric of accuracy. Accuracy was computed as the fraction of items in the test set for which the models' predictions were correct.

5.1 Questionnaires

A 5-point Likert scale was used in all questionnaires. Questions that ask for participants' rating on a bipolar 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. 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: easy, T_2: average, and T_3: difficult). 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 the Bonferroni correction was applied, and so all effects are reported at a .016 level of significance.

Table 2 shows the means and standard deviations for participants' assessment of the performed tasks. With respect to the assessment of the difficulty level it appears that there is a trend, with T_3 considered the most difficult. Friedman's ANOVA was applied to evaluate the significance of this variance. The results indicate that participants' perception of the task difficulty was significantly different ($\chi^2(3, N = 16) = 9.042$, $p < .05$). The post hoc tests show that the differences for the pairs T_1 & T_3 (Z = -2.434, $p < .016$) and T_2 & T_3 (Z = -2.683, $p < .016$) are statistically significant, but the same condition does not apply for T_1 & T_2.

This is further supported by participants' view of the retrieved results. The participants were asked to provide their assessments through the following questions: (i) "Overall, the results that were presented to you were: (Range: 1-5, Lower = Relevant - Higher = Irrelevant)", and (ii) "You feel satisfied with the retrieved results (Range: 1-5, Lower = Agree - Higher = Disagree) ". For the first question, the participants considered the retrieved results less relevant for T_3 (M=2.8125, SD=0.9106), compared to T_1 (M=2.0625, SD=0.8539) and T_2 (M=2.1875, SD=0.9811). Furthermore, they were less satisfied with the retrieved results in T_3 (M=2.8750, SD= 0.9574), than in T_1 (M=2.0625, SD= 0.9287) or T_2 (M=2.1250, SD= 0.9574). Table 2 also shows participant's assessment of the ambiguity, complexity and interest of the three tasks. Friedman's ANOVA test did not reveal a significant difference for any of the above aspects.

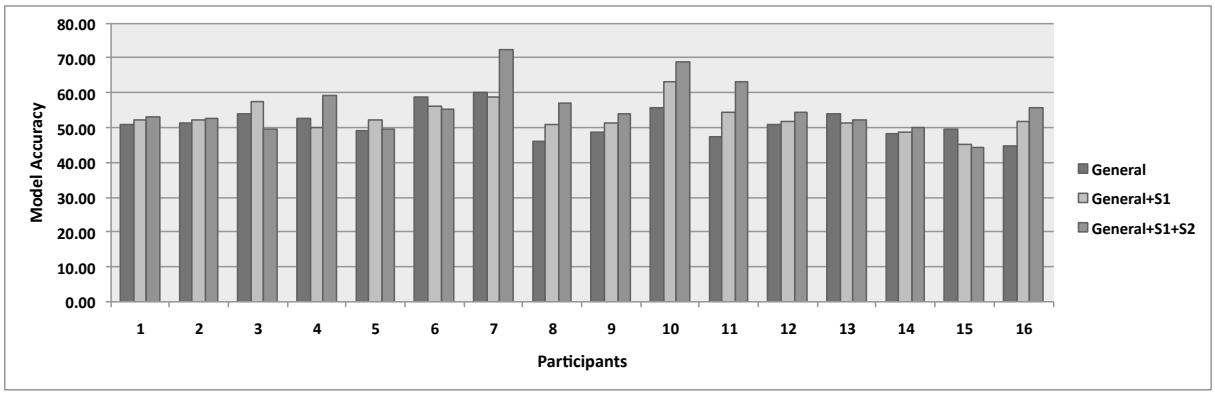

Figure 1: Results for models adapted using personalised data

5.2 Models

For each personalisation approach we present the performance of our models in terms of accuracy. The Dependent t-Test was applied, when possible, to determine if the difference between the experimental conditions is statistically significant. The baseline, which represents random choice, is set to 50%, since the class of a document can be either relevant or irrelevant.

5.2.1 Adaptation

The results of the first approach are shown in Figures 1, 2, and 3. Figure 1 illustrates the performance of three different classifiers, per participant: (i) a classifier trained exclusively on general data, (ii) a classifier trained using general data merged with the personalised dataset S_1, and (iii) a classifier trained using general data merged with the datasets S_1 and S_2. For every participant we tested these three combinations against the corresponding S_3. Only for this case, the subsets S_1-S_3 were not balanced. Therefore, the contribution of personalised data by each participant varied. The progression of the columns in Figure 1 suggests that, in most cases, an improvement was achieved by introducing personalised data to the training set, reaching classification rates that exceed 70%.

Figures 2 and 3 show the results of the same personalisation approach, as described above, with the exception that this time we used balanced sets of personalised data (we re-sampled the datasets S_1, S_2 and S_3 to ensure that each participant contributed the same number of instances). This was a necessary step to allow for testing the significance of the variation introduced in the models' performance. Figure 2 shows the performance of a classifier trained using the original set of general data, merged with an additional N instances of general data, and a classifier trained using the same general set of data, merged with an additional N^* instances of personalised data. On average, the second classifier (M=52.74, SD=3.31) performed slightly better than the first classifier (M=51.23, SD=4.33). Therefore, we can argue that by adding N number of personalised data we achieved a slightly better performance, compared to adding the same number of general data. However, the post-hoc tests did not reveal a significant difference.

Figure 3 illustrates the performance of the two classifiers after adding N+N general data, or N^*+N^* of personalised data, to the original training set and tested against the corresponding test set S_3, for each participant. The results show that the second classifier attained a significantly higher performance (M= 55.94, SD=6.62) than the first classifier (M=50.83, SD=4.34), $t(15)$=-3.848, $p < .01$. In this graph the enhancement of the model's performance, due to the integration of additional personalised data, is much more evident.

5.2.2 Weighted voting

The results of the second approach are presented in Figure 4. In this approach we used the general and the personalised data separately, to train two classifiers and combine their predictions using weighted voting. The graph illustrates the performance of the classifiers for different thresholds. Each line in the graph is a different weight combination, e.g., for w_{gen}=0.0 and w_{pers}=1.0 we see the progression of the performance, between thresholds 0.0 to 1.0. The graph indicates that, on average, the best performance was held by the classifier with combination of weights w_{gen}=0.3 and w_{pers}=0.7, for threshold t=0.3. This suggests that the voting scheme worked better when more emphasis was put on the personalised model. However, the contribution of the general model was equally important, to keep the classification rates optimal. When higher weights were given to the general model the performance dropped considerably, which supports further the positive effect of personalisation on the models' performance.

6. DISCUSSION & CONCLUSIONS

In this paper we explored two different approaches to personalising affective models that are capable of discriminating between two categories of documents: relevant and irrelevant. We devised an experimental setup that exposed our participants to search tasks of varying difficulty, which was achieved through the re-ranking of the return documents. This manipulation of task difficulty resulted in a much wider spectrum of affective reactions, thus making the accumulated affective data not only more authentic but also comparable to data gathered from previous studies. Our analysis also indicates that this variation was perceived by the participants, as it was found statistically significant.

For modelling relevance we extracted from facial expression data a set of features and classified them using Support Vector Machines. In the first approach we adapted a general model to the behavioural characteristics of a number

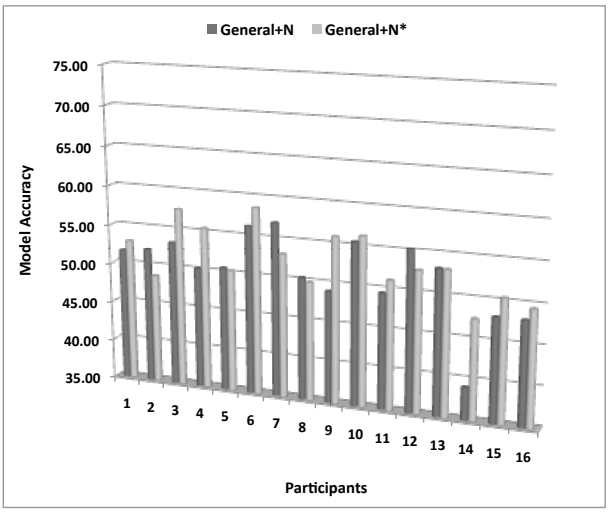

Figure 2: Performance of general model after adding N general or N* personalised data

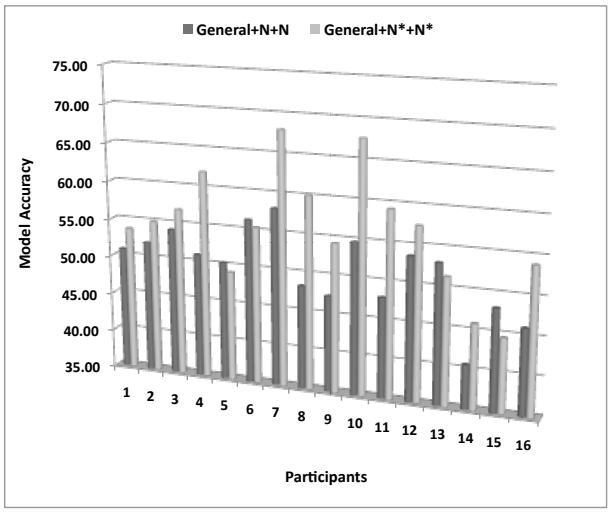

Figure 3: Performance of general model after adding N+N general or N*+N* personalised data

of participants, using personalised data, and established its performance against a model trained exclusively on general data. In the second approach we trained a general and a personalised model separately and combined their predictions using a combination of weighting schemes. We, finally, examined the effect of personalisation on the model's performance and tested it's significance.

One facet of affect recognition is developed here for the first time: the personalisation of affective models, trained on facial expression data, for the prediction of topical relevance. Our experimental evidence supports our first hypothesis, namely that by adapting a general affective model to a specific user we introduce a noticeable improvement in its discriminating ability. Our best performing model attained an accuracy of 72.52%, which is substantially better than the baseline or any other performance presented in [4].This difference was found to be highly statistically significant, which is an encouraging finding.

Using weighted voting we provided additional evidence in favour of accounting for the behavioural differences of users. Our analysis indicates that by fusing, on a decision level, the output of both general and personalised classifiers (with the emphasis on the latter) we can attain the optimal performance. Regarding our second hypothesis, we cannot suggest which approach was more effective, since our findings did not favour one method over the other. Clearly, there is more than one alternative to personalising user models, especially those built on affective data. Additional work is necessary before we determine if these two approaches perform equally well under different experimental conditions.

Finally, the evidence accumulated from both approaches suggests that personalisation works better for some users, than others. We speculate that the variation in the models' performance might be correlated with the ability of the participants to behave naturally and be expressive in a laboratory setting, as in their home environment. However, the choice of setting was a necessity, guided by the need to allow for comparability between data from previous studies.

In conclusion, we feel that the quality of our results is good enough to indicate that personalisation of affective feedback is a promising area of research and that it can 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. 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.

7. REFERENCES

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

[2] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–10, New York, NY, USA, 2006. ACM.

[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 *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 395–402, New York, NY, USA, 2008. ACM.

[4] I. Arapakis, I. Konstas, and M. Jose, J. Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In *Proceedings of the seventeen ACM international conference on Multimedia*, pages 461–470, NY, USA, 2009. ACM.

[5] 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 *Proceeding of the ACM International Conference on Image and Video Retrieval*, pages 1–8, NY, USA, 2009. ACM.

[6] R. Badi, S. Bae, J. M. Moore, K. Meintanis, A. Zacchi, H. Hsieh, F. Shipman, and C. C. Marshall.

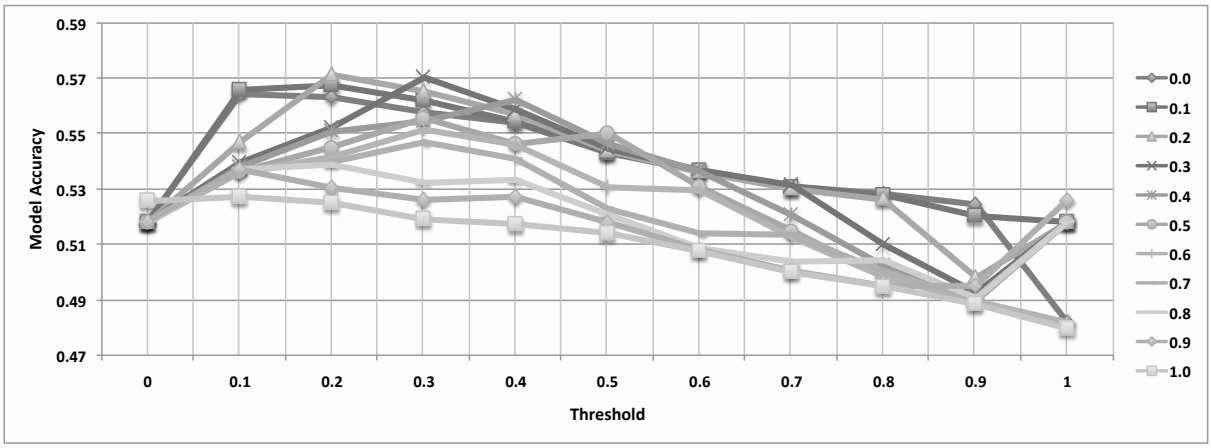

Figure 4: Results for the weighted voting (for different combinations of weights and threshold)

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.

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

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

[9] M. Daoud, L. Tamine-Lechani, and M. Boughanem. Learning user interests for a session-based personalized search. In *Proceedings of the second international symposium on Information interaction in context*, pages 57–64, New York, NY, USA, 2008. ACM.

[10] P. Ekman. *Facial Expressions*, chapter 16, pages 301–320. The Handbook of Cognition and Emotion. U.K.: John Wiley & Sons, Ltd, 1999.

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

[12] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161, New York, NY, USA, 2005. ACM.

[13] J. Koenemann and N. J. Belkin. 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*, pages 205–212, New York, NY, USA, 1996. ACM.

[14] F. Liu, C. Yu, and W. Meng. Personalized web search by mapping user queries to categories. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 558–565, New York, NY, USA, 2002. ACM.

[15] M. Morita and Y. Shinoda. Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 272–281, NY, USA, 1994. Springer-Verlag New York, Inc.

[16] D. W. Oard and J. Kim. Modeling information content using observable behavior, 2001.

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

[18] M. Pantic and L. J. M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, Sept. 2003.

[19] M. Pantic, N. Sebe, C. J. F., and T. Huang. Affective multimodal human-computer interaction. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 669–676, New York, NY, USA, 2005. ACM.

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

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

[22] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang. Authentic facial expression analysis. *Image Vision Comput.*, 25(12):1856–1863, 2007.

[23] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 449–456, New York, NY, USA, 2005. ACM.

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