Affect-Based Information Retrieval

Ioannis Arapakis

Department of Computing Science
Faculty of Computing Science, Mathematics and Statistics
University of Glasgow

UNIVERSITY
of
GLASGOW

Submitted for Degree of Doctor of Philosophy
at the University of Glasgow
February 2010

"There can be no knowledge without emotion. We may be aware of a truth, yet until we have felt its force, it is not ours. To the cognition of the brain must be added the experience of the soul."

Arnold Bennett (1867-1931)

Acknowledgements

This Ph.D is the product of personal research, under the supervision of Professor Joemon M. Jose of the Department of Computing Science of University of Glasgow. The conventions that dictate what a doctoral thesis should be like emphasise personal contribution to a large extent. That is, the doctoral thesis should be concerned with *my* work, the knowledge *I* have developed as a result of this work, and *my* interpretation of that knowledge. To my experience, research is not carried out only by individuals, nor do individuals develop new knowledge on their own. Rather science appears to be something that is woven from complex and fascinating social relationships between people. Discoveries are rarely made by a single person and insights tend to be conceived in situations where people communicate, share knowledge and ideas.

Writing an acknowledgments section that does justice to all the people who have contributed to this work is an impossible task. Nevertheless, I would still like to thank a number of people who's support made this three year long journey endurable and possible to complete:

Joemon M. Jose
for having enough faith in my skills to grand me this opportunity, and for being a great supervisor and friend.

Philip Gray and Keith van Rijsbergen
for your wisdom, advice and guidance. It has been a privilege to have the benefit of your counsel.

Peter Ingwersen and Alessandro Vinciarelli
for your constructive criticism and well-aimed suggestions for further improving my work.

the members of the Information Retrieval Group
for your kind support and interest in my research.

Ioannis Kompatsiaris and the members of MKLab

for a memorable collaboration and for participating so enthusiastically in my user experiments.

May Gallagher, Helen McNee, and the rest of the administration staff in the Department of Computing Science

you deserve a huge mention for keeping everything running smoothly.

Hideo Joho

for being an inspiration; your feedback and encouragement has been invaluable.

Ioannis Konstas, Vasilios Stathopoulos and Konstantinos Athanasakos

for your insightful comments and your active participation in my research.

Punitha Swamy

for regarding my setbacks with indulgence and for being a wonderful office-mate and friend.

Susanne Beate Oehler

for being a great flatmate, a good listener and a thoughtful friend.

Konstantinos Skoundras

for being my best mate, whether near or far; through good times and tough times.

my family

for your love, support and encouragement throughout life. This one's for you!

Abstract

One of the main challenges 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. 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. The formulated search queries, which are submitted to the retrieval systems to locate relevant items, produce poor results that do not address the users' information needs.

To deal with information need uncertainty IR systems have employed in the past a range of feedback techniques, which vary from explicit to implicit. The first category of feedback techniques necessitates the communication of explicit relevance judgments, in return for better query reformulations and recommendations of relevant results. However, the latter happens at the expense of users' cognitive resources and, furthermore, introduces an additional layer of complexity to the search process. On the other hand, implicit feedback techniques make inferences on what is relevant based on observations of user search behaviour. By doing so, they disengage users from the cognitive burden of document rating and relevance assessments. However, both categories of RF techniques determine topical relevance with respect to the cognitive and situational levels of interaction, failing to acknowledge the importance of emotions in cognition and decision making.

In this thesis I investigate the role of emotions in the information seeking process and develop affective feedback techniques for interactive IR. This novel feedback framework aims to aid the search process and facilitate a more natural and meaningful interaction. I develop affective models that determine topical relevance based on information gathered from various sensory channels, and enhance their performance using personalisation techniques. Furthermore, I present an operational video retrieval system that employs affective feedback to enrich user profiles and offers meaningful recommendations of unseen videos.

The use of affective feedback as a surrogate for the information need is formalised as the *Affective Model of Browsing*. This is a cognitive model that motivates the use of evidence extracted from the psycho-somatic mobilisation that occurs during cognitive appraisal. Finally, I address some of the ethical and privacy issues that arise from the social-emotional interaction between users and computer systems. This study involves questionnaire data gathered over three user studies, from 74 participants of different educational background, ethnicity and search experience. The results show that affective feedback is a promising area of research and it can improve many aspects of the information seeking process, such as indexing, ranking and recommendation. Eventually, it may be that relevance inferences obtained from affective models will provide a more robust and personalised form of feedback, which will allow us to deal more effectively with issues such as the semantic gap.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Outline

Preamble

In this chapter I highlight current challenges in Information Retrieval (IR) research related to the relevance feedback process and present existing methods and mechanisms designed to resolve these issues. Using the latter as reference I propose a novel approach that embodies the affective dimension of interaction. This new approach promises a more appropriate model, upon which to base a user-centred environment for search and retrieval. The background and motivation for the research are presented in subsequent chapters.

1.1 Introduction

One of the main challenges 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 noted, 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 (Belkin, 1980). The formulated search queries, which are used by the retrieval system to locate relevant items, produce results that do not address the user's information need.

To deal with information need uncertainty IR systems have employed in the past a range of feedback techniques. A key aspect of the feedback cycle is relevance assessments, which have become a popular practice in web searching activities and interactive IR. The value of relevance assessments lies in the progressive disambiguation of user's information need. This is achieved through the application of various Relevance Feedback (RF) techniques, which range from explicit (Koenemann and Belkin, 1996a; Rui and Huang, 2000) to implicit (Agichtein, Brill and Dumais, 2006; Badi, Bae, Moore, Meintanis, Zacchi, Hsieh, Shipman and Marshall, 2006).

Traditionally, explicit feedback methods necessitated the communication of user feedback through the explicit and intended indication of documents as relevant (positive feedback) or irrelevant (negative feedback). They are considered a robust approach to improving a retrieval system's performance, by offering better query reformulations and recommendations of relevant items (Koenemann and Belkin, 1996a). The notion of explicit feedback has been 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 (Belkin, Cool, Head, Jeng, Kelly, Lin, Lobash, Park, Savage-knepshield and Sikora, 2000).

The emergence of new trends in IR was followed by a shift of focus to the human factor. Implicit feedback techniques (Morita and Shinoda, 1994; Joachims, Granka, Pan, Hembrooke and Gay, 2005) were developed that collect interactional data, in an indirect and unobtrusive manner. By doing so, they disengage the users from the cognitive burden of document rating and relevance judgments. Observable behaviours such as reading time, saving, printing, selecting and referencing (Morita and Shinoda, 1994; Konstan, Miller, Maltz, Herlocker, Gordon and Riedl, 1997; Seo and Zhang, 2000) have been all treated as indicators of relevance, despite the lack of sufficient evidence to support their effectiveness (Nichols, 1997). Even though implicit feedback measures are still considered attractive and useful alternatives they are not always inherently so, simply because what can be observed does not necessarily correspond to the underlying intention. According to Kelly and Belkin (2002), implicit feedback measures that use interaction with the full content of documents can often be unreliable, difficult to measure, as well as interpret.

Moreover, existing feedback techniques determine topical relevance with respect to the cognitive and situational levels of interaction between the user and the system (Saracevic, 1975), failing to acknowledge the importance of intentions, motivations and feelings in cognition and decision-making (Damasio, 1994; Reeves and Nass, 1996; Pfister and Böhm, 2008). There is growing evidence that people naturally express emotion to machines and introduce a wide range of social norms and learned behaviours, which guide their interactions with, and attitudes towards, interactive systems and information items (Reeves and Nass, 1996; Vinciarelli, Suditu and Pantic, 2009).

A number of studies from the field of Library and Information Science (LIS) has provided evidence which shows that emotions can influence several aspects of the search process, such as search strategies (Nahl and Tenopir, 1996), performance (Wang, Hawk and Tenopir, 2000; Nahl, 1998b), and satisfaction (Nahl, 2004). Positive and negative emotions have been associated with satisfactory search results (Tenopir, Wang, Zhang, Simmons and Pollard, 2008), successful search completion (Bilal and Kirby, 2002), and interest in the process and documents (Kracker, 2002; Lopatovska and Mokros, 2008). Moreover, affective variables can play an important role in reading-related information behaviour, especially in the domain of everyday life (McKechnie, S. and Rothbauer, 2007). This suggests a need to understand RF more fully and re-examine it with respect to what occurs on the affective level of interaction.

However, the lack of consensus on what emotions are and how we can represent them has resulted in a vague discrimination of terms such as "emotion", "feeling", "mood", "attitude", and others. Many theorists have viewed emotions as response systems that coordinate actions, as affective feelings states, as interactions among different components (affective, cognitive, physiological, and emotional/expressive), or affective experiences, such as feelings or arousal and positive/negative valence. In this thesis, I use the terms "emotion" and "affect" interchangeably, and define the latter as the:

> *psycho-somatic mobilisation in response to the evaluation of an external or internal stimulus event, one that is perceived as relevant to the major concerns of the organism.*

According to Scherer (2005), examples of such events include natural phenomena, the behaviour of other people (external stimuli) or our own behaviour and memories (internal stimuli), which may be of importance to one's well-being. Both types of events can have a sudden onset and a very short duration, s or have a progressive onset and develop over time. Whatever the situation, emotions can be elicited due to changing conditions, accumulate and escalate. In the following section I identify the research objectives that have driven this work.

1.1.1 Research Objectives

Overall, this work is an exploration into a new territory, that of affect-based IR. The goal is to aid the information seeking process and improve the experience of a person searching, with a focus on the affective aspects of interaction. To achieve that goal, the following sub-goals were identified:

- Gather initial evidence on the role of emotions in the information seeking process.
- Associate users' affective responses to topical relevance of viewed results.
- Combine the accumulated evidence to train affective models that can effectively predict which information is relevant to the user.
- Explore different sensory channels, classification and personalisation techniques.
- Study the implications of the experimental conditions on users' search behaviour.
- Develop a platform (operational search environment) for evaluating affective feedback and comparing its performance to existing implicit feedback techniques.
- Investigate the ethical and privacy dimensions of affect-based interaction.

1.1.2 Research Hypothesis

In total, five evaluations were conducted as part of this thesis, which are presented in Chapters 4 to 8. All experiments involved human users that provided qualitative and quantitative data on search behaviour, via interaction logs and multi-modal input. In this context, the term multi-modality indicates the many different physically realised kinds of information, which can be exchanged between humans and machines. Bellow I present the research hypotheses that were investigated throughout the user experiments:

1. Chapter 4 serves as a starting point for the exploration of the role of emotions in the information seeking process and the impact of task difficulty on users' emotional behaviour. The research hypothesis examined here is:

   H_1: Users will experience the same affective states for different levels of task difficulty.

2. In Chapter 5 I evaluate the application of affective feedback, deriving from facial expression analysis, in a video search environment. The work presented here contributes to the exploration of the role of emotions in the search process, by highlighting some of the factors that can influence users' affective behaviour, and, furthermore, compares the performance of affective feedback to other implicit feedback indicators. The following research hypotheses are examined:

   H$_1$**:** Users' affective responses are consistent across different types of stimuli (search process, the viewed content).

   H$_2$**:** The integration of affective features, deriving from automatic facial expression analysis, in user profiling can improve the performance of a recommender system.

   H$_3$**:** Affective feedback can effectively complement existing feedback techniques, such as click-throughs.

3. In Chapter 6 I provide evidence of an associated relationship between affective feedback and topical relevance. I employ a range of sensory input, which range between facial expressions to physiological signals, to model user affective responses and predict the relevance of retrieved results. The latter occurs without the aid of explicit judgements. Overall, I examine the following research hypotheses:

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

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

4. In Chapter 7 I assess the effect of experimental conditions (naturalistic, semi-controlled, and controlled) on the affective models' performance. The size of the effect is determined by comparing the accuracy of models trained on sensory data from a semi-controlled experiment and models trained on data from a controlled experiment. The following research hypothesis is examined:

   H$_1$**:** Affective models trained on data gathered under semi-controlled experimental conditions will exhibit better performance compared to models trained on data gathered under controlled experimental conditions.

5. In Chapter 8 I investigate different ways of personalising affective models, trained on facial expression data gathered by many individuals. The objective is to determine whether the behavioural differences among users have an impact on the models' ability to discriminate between relevant and irrelevant documents. Using personalised data, I adapt the models to individuals and compare their performance to a general model. The research hypotheses examined here are the following:

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 compared to training separate models and applying information fusion on a decision level.

In the remainder of this chapter I provide an outline of this thesis.

1.2 Outline

This thesis is divided into the following chapters: Chapters 2 and 3 provide the background and motivate the work described in this thesis. In Chapters 4 to 8 I present a series of user experiments that investigate different aspects of the affective feedback framework, discussed in Chapter 2. In more detail, Chapter 4 explores the role of emotions in the information seeking process and the variation of affective response patterns for different settings of task difficulty. Chapter 5 evaluates the application of affective feedback in an operational video search environment, which applies real-time facial expression to discriminate between relevant and irrelevant results. Chapter 6 examines the use of affective feedback as an implicit indicator of relevance. Topical relevance is deduced implicitly, by measuring key physiological signals taken from the user during the evaluation of viewed results, unlike the surrogate methods discussed in the beginning of the chapter. Chapter 7 discusses the effects of the experimental conditions of the classifiers' performance. In Chapter 8 I present two different approaches to personalisation of affective models: one based on the adaptation of a general model, using personalised data, and one based on weighted voting using the predictions of a general and a personalised model. The main objective is to determine whether the behavioural differences among users have an impact on the models' ability to discriminate between relevant and irrelevant results. Chapter 9 presents a formalisation of the affective feedback framework, the Affective Model. The Affective Model is a cognitive model that motivates the use of evidence extracted from the psycho-somatic mobilisation, which occurs during the cognitive appraisal of an information item. Finally, in Chapter 10 I examine some of the ethical and privacy issues that arise from the social-emotional interaction, between human users and computer systems. The conclusions drawn from the user experiments and the thesis are discussed in Chapter 11, and avenues for future work are identified.

Chapter 2

Interactive Information Retrieval: An Overview

Preamble

This chapter provides the background for the research described in this thesis and creates a context, within which the work is situated. It presents the dominant frameworks in IR research: the *laboratory*, the *user-oriented* and the *cognitive* research approach. The chapter also critically reviews existing feedback techniques, highlights existing problems, and where appropriate it associates them to work presented in later chapters. It ends by introducing a new model of RF, one that accounts for the affective dimension of interaction and provides the theoretical grounds that support it.

2.1 The Laboratory Research Approach

Information Retrieval is one of the oldest disciplines in Information Science. It began developing four millennium ago as a product of humanity's need to organise and structure information for later retrieval and use. One of the original definitions of IR is the one coined by Mooers (1950), as cited by Savino and Sebastiani (1998):

> *Information retrieval is the name of the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him.*

IR has been traditionally concerned with the representation, storage, searching and locating of information that is relevant to a user requirement (Ingwersen, 1992). The main objective is to facilitate easier access to relevant information and promote the communication between the system and the user. Originally, IR followed the laboratory framework, which is a systems-oriented framework and suggests documents, search requests, their representation (queries), the database, and the matching of the latter two as the main line of research (Järvelin, 2007).

The laboratory approach does not consider the user as the initiator of the search topic. All topics are predefined and the evaluation is carried out using the standard IR metrics of precision & recall. The documents are regarded as information items that consist of text, written in some natural language, that lack of any structure other than phrasal, sentential, or paragraph. They have a known length and their content serves as the source of indexing features (Järvelin, 2007). Similarly to documents, the search requests (or topics) are also unstructured, natural language texts that represent information needs. They are associated with independent indexing features that derive from their content through natural language processing (NLP). These indexing features are usually words whose semantics help to describe the document's main themes and summarise its content (Baeza-yates and Ribeiro-Neto, 1999).

The indexing is carried out either manually or automatically. Manual indexing is performed by a group of experts in the domain of discourse, who examine the content of each document in the collection and annotate it, using representative terms (keywords) (Belew, 2001). According to Salton (1968), the manual indexing approach can offer improved levels of precision and recall, despite the cross-language issues. However, it is highly depended on the indexers' natural language processing skills and knowledge of the domain. On the other hand, automatic indexing is performed through the application of algorithms and morphological transformations (such as tokenisation, stop-word removal, stemming), and is, therefore, considered as less expensive and time-consuming. The comparison of these two methods has held vague and controversial results, which do not indicate clearly whether automatic indexing systems can attain a performance comparable to that of human indexing (Sensuse, 2004).

The independent features extracted during the indexing process are used to form queries, which are usually unstructured bag-of-words. Among their main attributes are the query length, specificity, type of content, search keys and query goals (Järvelin, 2007). User's uncertainty of the request/query is not taken into account since, as mentioned previously, the topics are pre-defined. Queries are used to produce ranked lists of document representations, in terms of score and (binary) relevance. Matching of the documents and queries is then performed, using the underlying retrieval model, and an evaluation of the top ranking documents is held using the standard criteria of precision and recall.

Precision (Belew, 2001) is the relationship between the number of retrieved relevant documents R, with respect to a query statement Q, and the number of documents D that have been retrieved based on it, i.e., R / D. However, if we want to evaluate the performance of a system in terms of retrieving every potentially relevant document in a collection, then we need to examine recall. Recall is defined as the number of relevant documents retrieved R, related to the total number of relevant documents in the collection C, i.e., R / C. An inverse relationship between precision and recall also exists but falls beyond the scope of this chapter.

As shown, the laboratory approach to IR has been predominately system-oriented and does not involve the participation of human users. It, therefore, allows better control over the experimental variables, as well consistency and comparability among the research findings. However, although these traditional goals have been very important it has become evident that IR is inherently an interactive process. This finding suggests that taking into account the interaction between the user and the IR components and processes is crucial for an effective system design and evaluation (Belkin, Cool, Stein and Thiel, 1995).

Most laboratory IR studies are constrained by the system's definition of needs and range of responses, which do not necessarily match those of users (Kuhlthau, 1991). Moreover, the evaluation metrics applied are limited, in many cases, to those of precision and recall. This rigid view of information seeking reduces a real-life process to a laboratory simulation that does not account for the inherent complexity and dynamics of physical actions, experienced affective feelings and cognitive thoughts, concerning processes and content.

In addition, the information needs are treated as static, suggesting that the information a user is looking for does not change during a search (Bates, 1989). Spink, Greisdorf and Bateman (1998) and Tang and Solomon (1998) argue that information needs should be regarded as transient, dynamic, mental constructs that develop over time, as a result of exposure to new information. Empirical evidence (Bruce, 1994) indicates that user's cognitive state can change during IR interaction, suggesting that information needs and relevance assessments are dynamic.

Finally, even though the techniques employed for the representation and handling of information items remain useful, they can be selected and controlled by human users instead of the system (Marchionini, 2004). Relevance judgments should be also applied by users, according to their needs and capabilities. Overall, the burden should be placed on the user, thus shifting from a system-optimising paradigm to a user-centred one, where the retrieval tasks are embedded in real-life activities and the system acts as the mediator that facilitates the interaction with the information sources.

2.2 The User-Oriented Research Approach

The emergence of the user-oriented framework brought a shift in the understanding of information, the nature of information needs (static vs dynamic), and the role of the user, the intermediary, and the IR system. According to Ingwersen (1992), the reasons that drove research towards this direction are threefold: (i) a need for a different perspective of the information seeking process, pointing to communicative and interactive IR models, (ii) a desire to develop a deeper insight into interactive processes and tasks carried out by intermediaries (human or system), and (iii) the theoretical drawbacks associated with the representation of meaning in IR.

This approach is mainly concerned with the behavioural dimensions of information seeking and information transfer. Unlike the laboratory approach, it focuses on the user and aims to improve the IR effectives through real-life, empirical investigations that take place within a social/organisational context (the centre and right-hand side of Figure 2.2). Here the emphasis is put on the individual user, the intermediary mechanism and the problem-solving processes that take place during the development and representation of information needs. Another prominent characteristic is the view of information. Unlike the view suggested by MacKay (1969), sequence of bits, symbols or signs are not regarded as information, but as data or potential information. They become information only when they are perceived as such by a cognitive agent, and understood within a non-scientific, emotional and cultural context (Ingwersen, 1992).

Furthermore, the user-oriented approach regards IR as a goal-driven, interactive process and attempts to model its dynamic and social character. The search requests are not merely predefined topics but rather real-life information requirements, which often take the form of ambiguous, ill-defined needs that evolve over time. Kuhlthau (1991) introduced a six-stage model of the Information Search Process (ISP), which describes the search from the beginning to the end, namely: (i) initiation, (ii) selection, (iii) exploration, (iv) formulation, (v) collection, and (vi) presentation.

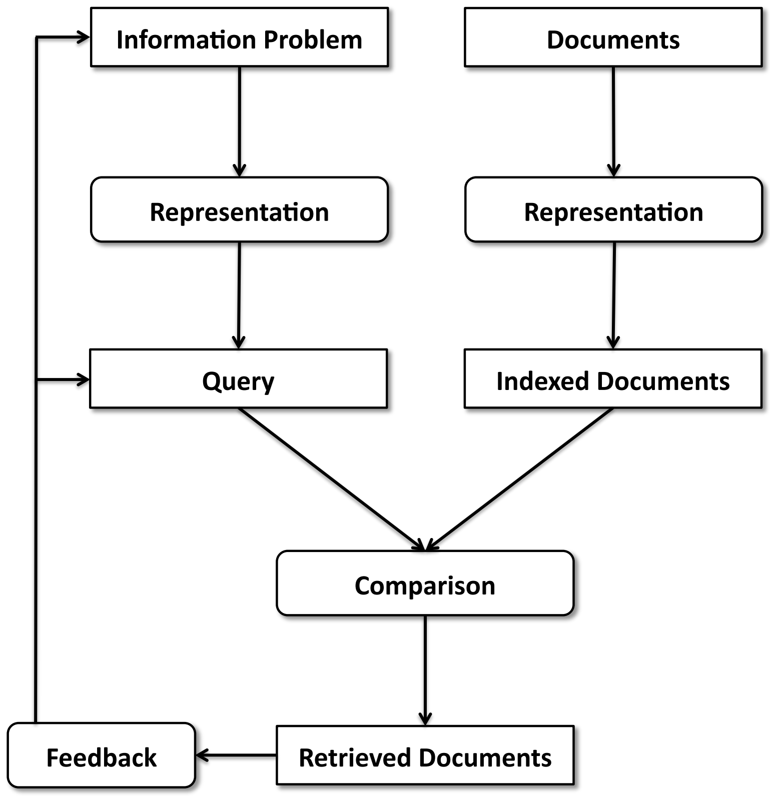

Figure 2.1: Information retrieval process (Croft, 1993)

According to Belkin (1980), an ISP is initiated when a user develops a need for information, deriving from an Anomalous State of Knowledge (ASK). This need for information is typically expressed in the form of a search query which is submitted to the IR system. Hiemstra (2000) argues that an IR system should support three basic aspects of the retrieval process, namely: (i) the representation of document content, (ii) the representation of the user's information need, and (iii) the comparison of the latter two representations. Figure

2.1 shows an example of an IR system. The rectangular boxes represent data items, while the round-shaped boxes represent processes.

The process of representing an information need as a query statement is often referred to as query formulation. This representation is compared across all documents within the corpus and a set of potentially relevant documents is returned. The user then interacts with the returned documents to determine their degree of relevance and indicate in an explicit, or implicit, manner which of them accommodate his/her need. Due to the exposure to new knowledge, or as an outcome of unsatisfactory results, the user might reformulate his/her initial query statement and re-submit it to the system. The quality of a query statement is dependable upon at least three variables, namely: (i) the understanding of the information need, (ii) knowledge of the document collection, and (iii) the mediator.

The realisation of this knowledge gap is the initiative step that motivates and drives an ISP. However, the information that one does not currently possess is information that one cannot describe in it's entirely (Campbell, 2000). A full description of something cannot be accomplished by anything other than the item itself, whether conceptual or physical. Hence, the user is only capable of describing what he knows at a given point, but not what he does not know yet and he is looking for. As a result, the user will formulate ill-defined queries that approximate his/her information need and, eventually, influence the quality of the retrieved results (according to the "quality-in quality-out" principle (Croft and Thompson, 1987)). Alternatively, a query statement that represents more accurately the desired information will produce better results.

Adequate knowledge of a collection can facilitate the formulation of better search queries. Simply put, a user who is familiar with the contents of a collection can structure query statements that have more discriminative power. This is a result of the knowledge of the collection's thematic structure. In addition, the mediator (or search environment) can play an important role in the query formulation process. Belkin (1980) suggests that the interaction between the user and the IR system occurs within a communication framework and underlines the difficulties the former is facing when articulating his information needs, due to cognitive or linguistic limitations.

IR systems that simply request keywords are considered suboptimal and often impose their own internal vocabulary to the user. Bennett (1972) suggests that the interface mechanisms built for IR systems should facilitate a more natural interaction, without adding to the cognitive efforts of the user. A study (Spink, 1994) on the selection of search term sources (taken from user question statements, user-interaction, term RF, and other) for expansion and reformulation revealed that the most effective source was terms taken from the user's written question statements. Based on this finding, the author suggests that IR interfaces should allow users to express their information needs using their natural language communication skills.

The user-oriented approach offered the IR community with a new insight into users' search behaviour, within an individual and a social context. It supported the management of information resources and through a series of real-life, empirical investigations it brought a new understanding of the formulation of information needs. Furthermore, it brought awareness to design issues related to the way IR systems accommodate users in fundamental IR processes, allow them to express their information requirements, and, finally, assist them in defining and solving the underlying problem (Ingwersen, 1992). In summary, the user-oriented approach shifted the focus to user-oriented activities of query formulation and reformulation, inspection and judgement of retrieved information items, as processes of central interest.

One of the major issues of the user-oriented approach is the development of an understanding of the nature of interaction itself and the transferring of this knowledge to the specification of systems' design. Without the existence of formal models of information behaviour and information interaction the experimental findings cannot offer insights valid for design and test purposes. Also, similarly to the laboratory approach, it did not grasp the full complexity of the information-related processes. Despite the empirical grounds that support its intuitive findings it lacks the theoretical grounds of the cognitive approach, which is discussed in the next section.

2.3 The Cognitive Research Approach

The cognitive paradigm introduced to IR an enriched theoretical context that encompasses and merges the two mainstream research approaches discussed earlier. In some aspects the analytic user-oriented approach can be seen as a predecessor of the cognitive turn in IR research. The cognitive view, however, focuses on the cognitive structures which are manifestations of human cognition, reflection or ideas (Ingwersen, 1996). These structures in the mind of each human recipient constitute his/her model of the world and consist of expectations, emotions, intuitions, intentionality and experiences. It is through these highly dynamic, changeable structures and their interaction that the individual perceives and understands the world.

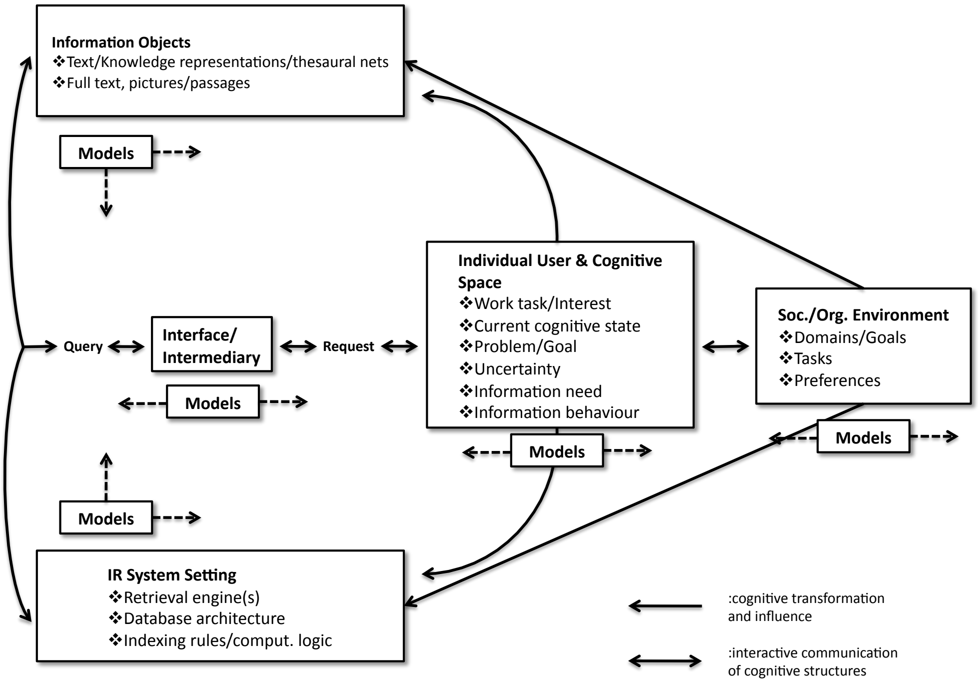

Figure 2.2: Cognitive communication model of information interaction in information science (Ingwersen, 1996)

In the context of IR these manifestations of human cognition can belong to a variety of cognitive origins, e.g., authors of texts, indexers, system designers, intermediary mechanism designers, and other. The cognitive communication model of information interaction in Figure 2.2 suggests some cognitive structures to be observed. With respect to the individual user it outlines the work task or interest that may lead to a perceived problem state (Ingwersen, 1994). The resulting cognitive situation may induce a certain amount of uncertainty and frustration to the user, as to how to reach his goal. This creates a need for external information (information need). The figure also suggests domains, goals, models, tasks and preferences (the social/organisation environment), as well as the intermediary and the IR system setting.

According to the cognitive approach, the role of the intermediary mechanism is to facilitate the communication between the user and the IR system, and even interfere at a very simplistic level. Ingwersen (1996) argues that the messages communicated from the system remain at a linguistic level until they are perceived by a human user and induce a change in his cognitive state. In the system setting a designer's cognitive structures may also be embedded in the IR system, under the form of specific database architectures, algorithms or models. In addition, the human indexers' cognitive structures can be represented by the index terms associated with each information item, while in the case of authors they can take the form of titles, captions, headings or cited works (Ingwersen, 1996).

Undoubtedly, the cognitive paradigm brought theoretical growth in IR research. It posed critical questions and suggested new concepts, such as the information need type, the work tasks and the semantic entities. Opposed to the laboratory approach, it aims to improve information access and support users in their information seeking activities by considering the interaction between the user and the IR system as a central event. It defines the system more broadly and evaluates how well the user, the intermediary mechanism and other entities interact under real-life conditions. Robertson and Hancock-Beaulieu (1992), as cited by Borlund (2000), states that

> *The conflict between laboratory and operational experiments is essentially a conflict between, on the one hand, control over experimental variables, observability, and repeatability, and on the other hand, realism.*

The cognitive approach examines phenomena and processes, such as information access, or the variation of precision & recall, by adopting an explanatory approach that focuses on user's cognitive states, besides documents and other knowledge structures (Järvelin, 2007). These states are linked to the perceived problem states, work tasks and information needs that exist within the cognitive space. Furthermore, it views the information need as a reflection of the ASK; as a dynamic as well as individual concept (Belkin, 1980). This suggests that, from the user's perspective, an information need is a personal, as well as individual perception, of a given information requirement (Borlund, 2000). Therefore, all relevance judgments should be made by the user, in relation to his personal information need, at the given time.

Finally, the interaction between the user and the IR system is seen in cognitive terms, without taking into account the importance of emotions in cognition and decision-making (Damasio, 1994; Pfister and Böhm, 2008). There is growing evidence which suggests that people naturally express emotion to machines along with a wide range of social norms and learned behaviours (Reeves and Nass, 1996; Picard, 2001; Picard, Wexelblat and Nass, 2002; Vinciarelli et al., 2009). While cognitive conceptions of the information need are necessary for research purposes, the affective dimension of user's problems, needs, and interactions with the information agents must be also examined, to allow the formulation of models that address a wider, holistic view of information access (Kuhlthau, 1991).

2.4 Relevance Feedback Re-examined

Relevance has been the subject of study by both philosophy and communication science. Several types and models of relevance have been formulated in the past, based on observations and empirical insights, in an attempt to explain our actions and connections to the real world. One contemporary philosophical interpretation of relevance is the one given by Schutz (1970), as cited by Saracevic (1975), who defined three types of relevance that interact dynamically and form what he called a *system of relevances*:

i) **Topical relevance:** perception of something being problematic.
ii) **Interpretational relevance:** involves the stock of knowledge at hand in grasping the meaning of that which is perceived.
iii) **Motivational relevance:** involves selection (the course of action to be adopted).

In the context of communication theory relevance is considered a "measure of the effectiveness of a contact between a source and a destination in a communication process" (Saracevic, 1975, p.325). This view of relevance is based on Shannon's (2001) model of communication, which consists of five elements, namely: (i) an information source that produces a message, (ii) a transmitter that encodes the message into a signal, (iii) a channel, adjusted to forward the signal, (iv) a receiver that decodes the signal to its original form, and (v) a destination for the message. According to this model, any form of communication consists of transmitting information from a source to a destination and can be viewed as a process on its own. Relevance can be seen as a measure of changes that occur during the communication of knowledge, represented as memories, collections, catalogs, files sentences, and other.

Despite its intuitive approach to relevance, the communication theory remains a relatively new field which stands deeply rooted within social structures and individual human behaviour. There is a range of socio-cultural and individual factors that are not accounted for in Shannon's theorem, which oversimplifies the process of communication. According to Saracevic (1975)

> *Knowledge, information, communication, information systems-all are imbedded in, all reflect some system of human values-ethical, social, philosophical, political, religious and/or legal values. Therefore, when considering relevance, one may also involve aspects of the environments, realities and values.*

While taking into account the theoretical definitions of relevance that were formulated by philosophy and communication, Saracevic (1996) proposed a set of general characteristics. This conception of relevance suggests that we may need to consider several types of relevance and their interdependencies, as different elements come into play. More specifically, the attributes consist of the following:

i) **Relation:** Relevance indicates a relation. Different manifestations of relevance encompass different relations.
ii) **Intention:** Objectives, roles, expectations and motivation.
iii) **Context:** The matter at hand - the context from which the expression of relevance arises.
iv) **Inference:** Relevance can be intuitively understood as a human built-in mechanism that is based on cognition. It involves the assessment of a relation and its effectiveness, in relation to an intention and a context.
v) **Interaction:** a dynamic and interactive process, where interpretations of other attributes may be affected by changes in the cognitive space of the individual.

From the early years of information science relevance was a key notion and one of its main objectives, that is to facilitate easy access to *relevant* information. Nowadays, It is regarded as a process which is intertwined with several cognitive functions, such as communication, information seeking, information assessment, reflection, and other. This shift of focus to the notion of relevance was motivated by problems that arise due to the substantial increase in the number of scientific publications and the breakdown of boundaries between subjects. This caused a concern regarding the classification and indexing of information, and established relevance as an assessment criterion.

In the domain of IR, relevance has been viewed in a similar manner. The term relevance refers to *relevance feedback*, the process whereby an information request is modified automatically by the system. The refinement of the request is based on system feedback, e.g., the returned ranked list of information items (graphics, documents, etc.) (Baeza-Yates and Navarro, 1996; Koenemann and Belkin, 1996b). Those items identified by the user as relevant are used to adjust the weighting of the terms in the search query, which is then resubmitted to the IR system (Belkin and Croft, 1992). The refined query produces new results which, in theory, are semantically and contextually closer to the user's information needs. Relevance feedback can be, therefore, described as an iterative and interactive process that facilitates the retrieval of items similar to those judged relevant by the user (Figure 2.3).

In the laboratory approach, IR representations, algorithms and models were, and still are, evaluated with respect to relevance. However, this type of evaluation does not involve

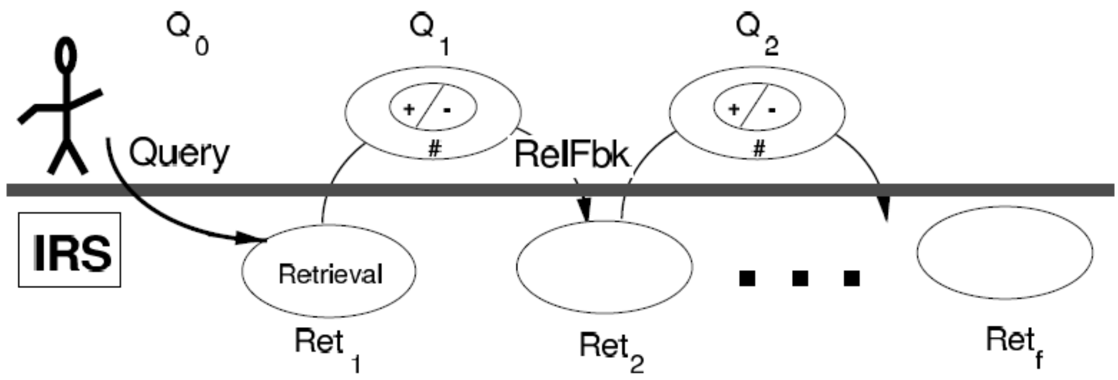

Figure 2.3: The cycle of relevance feedback (Belew, 2001)

the participation of human users, which, as discussed in Section §2.1, is its main drawback. Moreover, without the involvement of an intermediary mechanism to regulate the communication between the system and the user, the former lacks the necessary input to make accurate decisions regarding query re-structuring and the use term weights (Ingwersen, 1992).

On the other hand, the user-oriented approach takes into consideration the effects of interaction and RF on user's cognitive structures and problem space. The information need and the search query are no longer considered constants but are treated as variables, which evolve over time, while the role of test collections and comparative methods becomes operational compared to the laboratory framework. This research approach creates a firm bond between the IR process and the users, who become the assessors of relevance.

From a cognitive point of view RF can be described as an interactive task of object recognition (Belew, 2001), where the most intelligent component is the user, despite his/her fragmented knowledge of the problem space. The object to be identified is the user's mental representation of the document that meets the desired criteria and can satisfy the developing information need. In other words, the cognitive notion of relevance suggests the dynamic and interactive establishment of a relation, by means of inference, towards a context. The best way to establish this relation is by providing the user with conceptual support as early as possible during the ISP, to aid in defining and formulating his need and, eventually, receive potentially relevant contexts (Ingwersen, 1992).

Although this approach introduced an improvement to the IR metrics of precision and recall (Christiansen and Lee, 2006), placing the burden of relevance judgements on the user (instead of computer heuristics) remains a cognitively demanding and time-consuming task. The cognitive resources that need to be allocated depend on the level of understanding of the information need, the exhaustiveness of the assessment and the employed RF technique.

Another important issue is user consistency (Zhou and Huang, 2001). RF is non-metric data, which suggests that even though users consider it relatively simple to tell between two items which one is more relevant to their information need, it is difficult to determine the degree of relevance and even harder to keep the judgements consistent. Salton and Buckley (1990), as cited by Belew (2001), argue that this phenomenon of inconsistency is not observed only during relevance assessments, but in any kind of task that involves the evaluation of two or more items and the quantification of the likeness on a metric scale.

Regardless of the point of view, whether philosophical or scientific, all approaches to relevance have a common denominator: relevance denotes a relation and different manifestations of relevance encapsulate different relations. However, there is a lack of consensus on the taxonomy of these manifestations, even though many of the types and kinds share similar aspects. Saracevic (1996) proposes the following types of relevance, within the context of IR and information science:

i) **System or algorithmic relevance:** the relation between a query statement and the information items in a system as retrieved, or not, by a given algorithm. Each system has its own way of representing, organising and matching these items to a query.
ii) **Topical relevance:** the relation between the subject or topic expressed by a query and the subject or topic covered by the retrieved information items. Aboutness is the criterion by which topicality is inferred.
iii) **Cognitive relevance:** the relation between the state of knowledge and the information need of a user and the retrieved information items.
iv) **Situational relevance:** the relation between the situation, task, or problem at hand, and the retrieved information items.
v) **Motivational or affective relevance:** the relation between the intention, goals and motivations of a user and the retrieved information items.

To conclude, RF is regarded as an iterative and interactive process that can aid in the disambiguation of a user's information need, as well as improve the effectiveness of information searching (Harman, 1992; Salton and Buckley, 1997). There can be different types of relevance, with each type reflecting one of the relations presented above. In addition, relevance feedback holds three main advantages over other query reformulation strategies (Baeza-yates and Ribeiro-Neto, 1999): (i) it shields the user from the details of the query reformulation process, (ii) it introduces the ISP as a sequence of smaller and uncomplicated steps, and (iii) it offers a controlled approach to query modification, designed to promote the relevant index terms and de-emphasise the irrelevant ones. To deal with information need uncertainty and tailor the search criteria to the user needs IR systems have employed a range of RF techniques, which vary from explicit to implicit. In the forthcoming sections I will present both types of feedback techniques that have been applied in the domain of interactive IR.

2.4.1 Explicit Feedback Techniques

In recent years RF has become a popular practice in web searching activities, and it is viewed as an effective method in dealing with difficulties that arise during query reformulation. For an average user, expressing an information need in the form of a logically structured query can be often a challenging task, especially for those who do not possess adequate experience or training. The RF cycle can aid the user by suggesting improved query statements and recommend unseen items, based on their early assessments. The type of user feedback that is obtained through the explicit and intended indication of information items as relevant (positive feedback) or irrelevant (negative feedback) is known as explicit RF.

Explicit feedback is a robust method for improving a system's overall retrieval performance and providing better query reformulations (Koenemann and Belkin, 1996a), at the expense of user's cognitive resources (Belkin et al., 2000). A number of studies (Koenemann and Belkin, 1996a; Beaulieu, 1997; Konstan et al., 1997; Belkin et al., 2000) provided evidence which indicate that explicit RF based techniques are generally desirable, even though they are scarcely applied during a search process (Beaulieu, 1997; Belkin, Cool, Kelly, Lin, Park, Perez-Carballo and Sikora, 2001). It is evident that there is a controversy among the features users claim to prefer and the ones they actually apply, when interacting with IR systems. Furthermore, as the level of search complexity increases the user's cognitive resources stretch even thinner making cognitive tasks, such as the assessment of documents, non-trivial (Beaulieu, 1997; Belkin et al., 2001).

White (2004), citing Aalbersberg (1992), argues that explicit feedback mechanisms haven't met significant commercial success, despite the apparent benefits they introduce, for two reasons: (i) the high computational cost for the application of the RF algorithms, and (ii) the unfriendliness of RF interfaces. However, recent advances in the hardware and software engineering have allowed the development of RF techniques that exhibit increased usability and effectiveness. Despite these improvements explicit RF techniques still 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 are reduced, turning relevance assessment into a cognitively demanding task (Belkin et al., 2000). White (2004) argues that this problem escalates during the early stages of the ISP, where the user is expected to evaluate a number of retrieved documents. Often, due to the variability in the size of the documents and the complexity of the topics, the users may need to allocate most of their time in these assessments, despite the lack of time or willingness to do so. In addition, their judgments can be limited to the documents that were retrieved by the system, which makes the effectiveness of a RF technique in inferring an information need dependable on the degree of recall.

The recent increase in the literature covering interaction-related aspects of IR has raised awareness on a number of design and implementation issues, some of which are specific to the ISP and the role of feedback techniques. Despite their intuitive and straightforward character, researchers have begun questioning the level of support these techniques offer (Bates, 1990). Buckley, Salton and Allan (1994) argue that the design of existing RF systems does not provide adequate information to support the effective operation of the underlying query re-formulation heuristics and algorithms, discouraging users from applying relevance assessments to the viewed items. In addition, the indirect control over the query formulation, through the communication of RF, requires an understanding of the underlying semantics, concepts and the making of meaningful decisions. This may not always facilitate an easier access to relevant information or promote an effective communication between the system and the user (Beaulieu and Jones, 1998). Increasing the availability of information and the range of decisions will most likely result in an increase of the user's cognitive load.

In this section I discussed the advantages and disadvantages of explicit RF techniques. As Bates (1990) suggests, users naturally expect to retain a level of control over the search process and the retrieval parameters, yet the explicit feedback techniques tend to counterbalance this control with the cognitive burden and the additional complexity they introduce. In the following section I present another category of RF techniques, which determine user interests and information needs by capturing interaction and contextual data in an indirect manner.

2.4.2 Implicit Feedback Techniques

The feedback techniques that fall under this category collect information on user search behaviour in an unobtrusive and sophisticated manner. By doing so, they disengage the user from the cognitive burden of document rating and relevance judgments. Information-seeking activities such as reading time, saving, printing, selecting and referencing (Morita and Shinoda, 1994; Joachims, Freitag and Mitchell, 1997; Konstan et al., 1997; Billsus and Pazzani, 1999; Seo and Zhang, 2000) have been all treated as indicators of relevance, despite the lack of sufficient grounds to support their effectiveness (Nichols, 1997). Among the main advantages of implicit feedback techniques is that they remove the cost to the user of providing feedback (Kelly and Teevan, 2003). In addition, it is relatively easy to accumulate large quantities of implicit data without the intentional participation of the users. Considering the latter, implicit feedback techniques become attractive alternatives despite their lack of accuracy (Nichols, 1997).

The first known taxonomy of implicit feedback indicators was introduced by Nichols (1997), in a study of the costs and benefits of implicit ratings for an information filtering application. The categories that appear in this taxonomy involve user activities observed during the interaction with the intermediary mechanism. Nichols argues that the implicit ratings have potential, but the limited availability of evidence does not offer sufficient ground truth in favour of their effectiveness.

Claypool, Le, Wased and Brown (2001) carried a similar assessment of implicit feedback indicators and proposed a set of observable behaviours that can be used as criteria for detecting user interest. In this study, more than 70 undergraduate students were asked to perform a random search on the web using a custom-made browser that captured interaction data, and among them several types of implicit interest indicators. The implicit indicators consisted of mouse clicks, mouse movements, scrolling and elapsed time. The participants were asked to explicitly rate each page they visited and these ratings were used as the ground truth. The findings suggest a strong correlation between certain implicit feedback indicators and users' explicit ratings. However, due to the lack of predefined search tasks and the uncontrolled nature of the study, the results are considered rather generic.

Oard and Kim (2001) proposed a taxonomy of observable user actions, very similar to that of Nichols's (1997). The actions are mapped using a two-axis categorisation system: behaviour and minimum scope. As seen in Table 2.1 the vertical axis is the primary organising principle (the purpose of the behaviour), while the horizontal axis is the secondary categorisation (the minimum scope of the item under use). The vertical axis suggests four different categories of behaviours that can occur while the user is interacting with the intermediary mechanism, namely: (i) examining a document, (ii) retaining a document, (iii) referring to a document, and (iv) annotating a document. The horizontal axis suggests different levels of scope of the associated behaviour, which can vary from small (viewing a paragraph or snippet) to large (viewing an entire document).

Kelly and Teevan (2003) extended the feedback indicators taxonomy originally developed by Oard and Kim (2001) by adding a fifth behaviour category, *Create*. The actions that fall in this category relate to user behaviours that are exhibited during the creation of information, e.g., the writing of a document. Although this taxonomy does not exhaust all observable behaviours it includes a sufficient range of interest indicators. White (2004) argues that only the *Examine* and *Retain* categories are suitable for classifying user behaviour since the rest of the actions require a higher level of control over the document, which is rarely made available to the user.

		Minimum Scope		
		Segment	**Object**	**Class**
Behaviour Category	**Examine**	View, Listen	Select	
	Retain	Print	Bookmark, Save, Delete, Purchase	Subscribe
	Reference	Copy-and-paste	Forward, Reply, Link, Cite	
	Annotate	Mark up	Rate, Publish	Organise

Table 2.1: Observable behaviours in IR interaction (Oard and Kim, 2001)

The application of implicit feedback techniques has caused some controversy in the IR community. Despite their apparent advantages they suffer from low accuracy and do not account for the individual and behavioural characteristics of each user. It is assumed that users will exhibit, in their majority, the same stereotypical behaviour when exposed to topically relevant information. One such example is the amount of time a user will spend assessing a document (reading time), which has been the subject of debate in the past (Konstan et al., 1997; Billsus and Pazzani, 1999; Seo and Zhang, 2000; White, Ruthven and Jose, 2002). Reading time is generally regarded as an ambiguous and unreliable interest indicator, primarily because it does not account for other aspects that might be brought out or emphasised, such as the nature of the task, the topic at hand and the individual characteristics of the user (Kelly and Belkin, 2001; Kelly and Belkin, 2002). Furthermore, the behaviour of the user, with respect to reading time, can vary significantly at different stages of the ISP. A study on newsgroup articles by Morita and Shinoda (1994) provided evidence that suggest a correlation between the time spent on reading an article and user preference. The participants of the study exhibited a tendency to spend more time reading articles that were rated later on as interesting and less time reading articles that arouse little, or no interest at all.

Kelly and Belkin (2002) and Kelly and Cool (2002) came up with contradictory findings, which demonstrate that when users deal with unknown topics, for which they do not possess domain-specific knowledge, they will likely display in their browser windows both relevant and irrelevant documents for an equal amount of time. As their familiarity with and knowledge of the domain increases so does their ability to discriminate between the two categories of documents. Consequently, search efficacy will be inversely analogous to reading time, which will progressively decrease.

In a naturalistic study Kelly (2004) found no direct relationship between display time and document usefulness. The display times were found to be significantly related to the information-seeking context in different ways, for different users, and varied according to task, topic, and topic familiarity. Furthermore, retention did not always prove a good indicator of document preference. Kelly (2004) suggests the modelling of user information behaviour as an approach to inferring document relevance, and that such approach should be personal rather than general.

Some encouraging examples of the use of implicit feedback can be found in studies that apply user modelling (Jameson, Schäfer, Weis, Berthold and Weyrath, 1999; Schwarzkopf, 2004; Badi et al., 2006; Tamine, Boughanem and Zemirli, 2006). In most cases, the necessary information to build the model is gathered by the intermediary mechanism, i.e., the interface environment. This approach offers the advantage of adapting the model according to the encountered search conditions and the person by automatically learning user behaviour (Agichtein, Brill, Dumais and Ragno, 2006). Overall, the evaluations of user modelling techniques indicate an improvement in the detection of document relevance over other techniques. Nevertheless, even this methodology does not address the full complexity and diversity of the ISP. Additional work is necessary before developing models that exhibit sufficient robustness towards individual, situational and contextual variation (Badi et al., 2006).

From the work presented in this section it is made evident that several reliability issues arise when attempting to infer RF from observable search behaviours. This is 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. Kelly and Belkin (2002) argue that implicit feedback measures that use interaction with the full content of documents can often be unreliable, difficult to measure and interpret. Relevance feedback should be understood within the larger context of the users' goals, intentions, motivations and feelings.

2.4.3 The Affective Dimension of Relevance Feedback

According to Saracevic (1996), IR interactions occur in several overlapping layers that involve different components and processes. On the user side the nature of these processes can be psycho-physiological or cognitive, while on the system side it can be physical, symbolic or algorithmic. The interaction between user and system takes place on the *surface* level, with the help of an intermediary mechanism. Within the context of this dialogue the user engages in a number of processes (browsing, navigating, assessing, visualising, inferring relevance, etc.) and interacts with information resources on a cognitive, situational and affective level. The system responds to the user's requests based on algorithms and procedures provided. On the *cognitive level* the information resources are considered as cognitive structures, which the user interprets, understands, assimilates and processes cognitively. On the *situational level* the user deals with a practical problem, e.g., a knowledge gap, that results in the development of an information need and the associated query. As additional dimensions may be brought in or de-emphasised, the information need and the query can develop and change.

As discussed in Sections §2.4.1 and §2.4.2, by analysing explicit and implicit feedback we can determine topical relevance, offer improved query reformulations and tailor the search criteria to the user needs. However, existing feedback techniques determine topical relevance in respect to the cognitive and situational levels of interaction, failing to acknowledge the importance of intentions, motivations and feelings in cognition and decision-making (Damasio, 1994; Pfister and Böhm, 2008).

There is growing evidence that people naturally express emotion to machines and introduce a wide range of social norms and learned behaviours that guide their interactions with, and attitudes toward, systems and information items (Reeves and Nass, 1996; Picard, 2001; Picard et al., 2002; Vinciarelli et al., 2009). Therefore, there is a need to reconsider RF with respect to the affective level of interaction and, furthermore, study the *individual differences* in affective behaviour, since the latter consist one of the most prominent features in relevance inferences (Saracevic, 2007).

In an earlier study, Kuhlthau (1991) proposed a six-stage model for the 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 dimensions 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. Additional findings by Nahl and Tenopir (1996) and Nahl (1998a; 2004) suggest the correlation of affective, cognitive and physical behaviours. More in specific, Nahl (1998a) found that the affective component of information search behaviour can regulate cognitive processing through a hierarchical organisation of goals, which is prescribed by both individual and cultural elements.

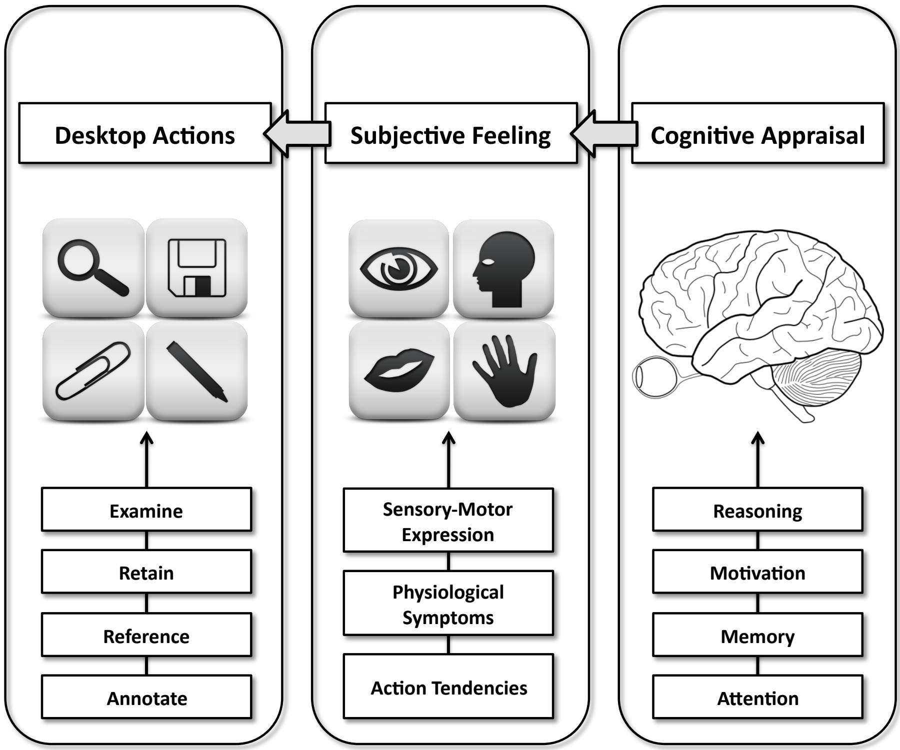

Figure 2.4: Levels of interaction

A number of studies from the field of Library and Information Science (LIS) have provided evidence that highlight the role of affect in several aspects of the information seeking process, such as search strategies (Nahl and Tenopir, 1996), motivation (Nahl, 2004), performance (Nahl, 1998b; Wang et al., 2000; Nahl, 2004; Nahl, 2005; Kim, 2008; Tenopir et al., 2008) and satisfaction (Bilal and Kirby, 2002; Nahl, 2004). Positive and negative emotions have been associated with satisfactory search results (Tenopir et al., 2008), successful search completion (Bilal and Kirby, 2002), interest in the process and documents (Kracker, 2002; Lopatovska and Mokros, 2008), as well as content design and aesthetics (Lavie and Tractinsky, 2004). McKechnie et al. (2007) suggest that the affective variables can play an important role in reading-related information behaviour, especially in the domain of everyday life. Information processing, which occurs during the appraisal process of a goal, an event, or an item, can result in a series of changes in the user's cognitive and affective states (Scherer, 2001).

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

Furthermore, these unconscious changes follow cognitive appraisal and precede any actions taken by the user, as shown in Figure 2.4. This suggests a number of things. Firstly, user affective responses cannot be easily faked due to their involuntary character. For example, variations in the levels of body perspiration or heart rate, like many other bodily functions, are controlled by the Sympathetic Nervous System (SNS), which is part of the Autonomic Nervous System (ANS). ANS acts as a control system that functions bellow the level of consciousness, thus most of its actions are involuntary. On the contrary, user actions that occur on an interface level can be potentially faked and mislead the feedback model. Secondly, affective responses are considered more spontaneous and authentic, due to their direct relation to the outcome of the appraisal process. Finally, they can be captured using a variety of sensory channels that outnumber the limited range of implicit, and often unreliable, feedback indicators.

The modelling and integration of affective features in the feedback cycle could allow IR systems to facilitate a more natural and meaningful interaction. Furthermore, it could improve the quality of the query suggestions, and, potentially, influence other facets of the information seeking process, such as indexing (Chan and Jones, 2005; Hanjalic and Xu, 2005), ranking (Soleymani, Chanel, Kierkels and Pun, 2008) and recommendation (Arapakis, Moshfeghi, Joho, Ren, Hannah and Jose, 2009). Eventually, it may be that relevance inferences obtained from this kind of models will provide a more robust and personalised form of feedback, which will allow us to deal more effectively with problems such as the semantic gap.

2.5 Summary

In this chapter I have described the background and motivation behind the work presented in this thesis. It is made apparent that existing RF techniques suffer from a number of problems that constitute their application either cognitive-expensive or unreliable. There is a need for techniques that will facilitate, in an unobtrusive and intelligent manner, a more effective information access, by reducing the number of search decisions and actions the users must make. A new concept of RF is proposed, that offers an alternative approach to modelling user behaviour and representing the developing information needs. The modelling techniques described in this thesis exploit affective information, channelled through various modalities, to make decisions that determine the topical relevance of information items. In the following chapter I discuss theories and definitions of emotions, present methods of measurement and, finally, review existing work in IR research.

Chapter 3

Emotion Research

Preamble

Online IR is a process that is influenced by and results in changes of the user's affective states. This chapter reviews literature on the theories of emotions, methods for studying emotions, and their role in human-computer interaction. Finally, it presents existing user-centred and content-centred studies of emotions in IR.

3.1 Introduction

Evidence from recent neurological studies underlines the importance of emotions in human cognition and perception (Picard, 2001). Emotions play an essential role in social interactions (Scherer, 2003; Russell, Bachorowski and Fernandez-Dols, 2003; Sander, Grandjean, Pourtois, Schwartz, Seghier, Scherer and Vuilleumier, 2005), perform important regulatory and utilitarian functions within human body and brain (e.g., flight or fight), and facilitate rational decision making and perception (Damasio, 1994). In recent years there has been an increased interest in emotion theory in disciplines such as LIS (Nahl and Bilal, 2007) and IR (Arapakis, Jose and Gray, 2008; Lopatovska and Cool, 2008), besides human-computer interaction (Picard, 1997; Julien, McKechnie and Hart, 2005). Some progress has been made in developing "affective systems" that are capable of recognising and responding appropriately to human emotions, ultimately making human-computer interaction experiences more effective and pleasurable.

Emotion theory is an amalgam of different approaches to the study of emotion, mainly based on cognitive psychology, with contributions from physiological psychology, philosophy, economics, engineering, and computer science. Psychology has the longest history of emotion research and has developed a solid framework for studying emotions. This chapter reviews definitions and theories of emotions, methods for studying emotions and, finally, provides examples of emotion research in MultiModal Human-Computer Interaction (MMHCI) and IR.

3.2 Theories and Definitions

Despite the long history of inquiry into the nature of emotion there is an apparent lack of consensus and uniformity within the scientific community on what emotions are and how we can represent them. Kleinginna and Kleinginna (2005) collected more than 90 definitions of emotions (including 9 sceptical statements). Many theorists have viewed emotions as response systems that coordinate actions, as affective feelings states, as interactions among different components (affective, cognitive, physiological, and emotional/expressive), or affective experiences, such as feelings or arousal and positive/negative valence. Moreover, the indiscriminate application of the term "emotion" has led to the vague differentiation between the terms "emotion", "feeling", "mood", "attitude", and others.

Among the classical theories of emotion there are two predominant approaches. The first invokes appraisal as an evaluative or interpretive function that is critical for eliciting emotion, and attempts to explain the subjective manifestations of the emotional experience in terms of cognition. In emotion theory, cognition encapsulates perception, attention, evaluation, decision-making, memory, and other components. The cognitive theories of emotion suggest that appraisal can be conscious or unconscious, intentional or unintentional, and take the form of a judgement or a thought (Folkman, Lazarus, Gruen and DeLongis, 1986). It denotes the evaluation of a particular encounter with the environment or a situation, in terms of its relevance to one's goals or well-being (Lazarus, 1984). Frijda (1993) argues that appraisal is the means by which we extract meaning from, and react to, affectively important events and make sense of the world. By producing meaning, an appraisal sets the occasion for an emotional response whose purpose is to prepare cognition and facilitate a way of coping with the situation at hand (Lazarus, 1984).

Cognitive appraisal can be further divided into primary and secondary. In primary appraisal an organism evaluates a situation with respect to what is at stake (the criteria of this evaluation vary between values, commitments, beliefs, etc.), while in secondary appraisal the organism evaluates what can be done to prevent harm or improve the probability for benefit. Scherer (2005) suggests a classification of the appraisal mechanism between intrinsic and extrinsic. The former type of appraisal refers to the evaluation of an event, object, or person, independently of the current needs, desires and goals of an organism, while the extrinsic to the type of evaluation that considers all the latter.

The componential theory of emotion (Scherer, 2005) is another example of the modern cognitive theory that views emotion as a synchronisation of many different bodily, perceptual and cognitive processes, in response to the evaluation of an external or internal stimulus event (Figure 3.1) as relevant to major concerns of the organism (Scherer, 2005). Such examples of events include natural phenomena, the behaviour of other people (external stimuli) or one's own behaviour or memories (internal stimuli), which may be of importance to one's well-being. Both types of events can have a sudden onset and a very short duration or have a progressive onset and develop over time. Whatever the situation, emotions can be elicited due to changing conditions, accumulate and escalate.

The second approach is concerned with the cognitive function of emotions and emphasises the somatic factors in the perception of emotional expressions (Zajonc, 1984). Theorists that follow this tradition view emotions as causing cognitive changes and argue that their main function is to direct attention to relevant aspects of the environment, due to the in-built action tendencies for altering that environment (Ekman, 1984; Frijda, Manstead and Bem, 2000). From this perspective, attention, among many other cognitive processes, is always motivated and constrained by the experienced emotional state.

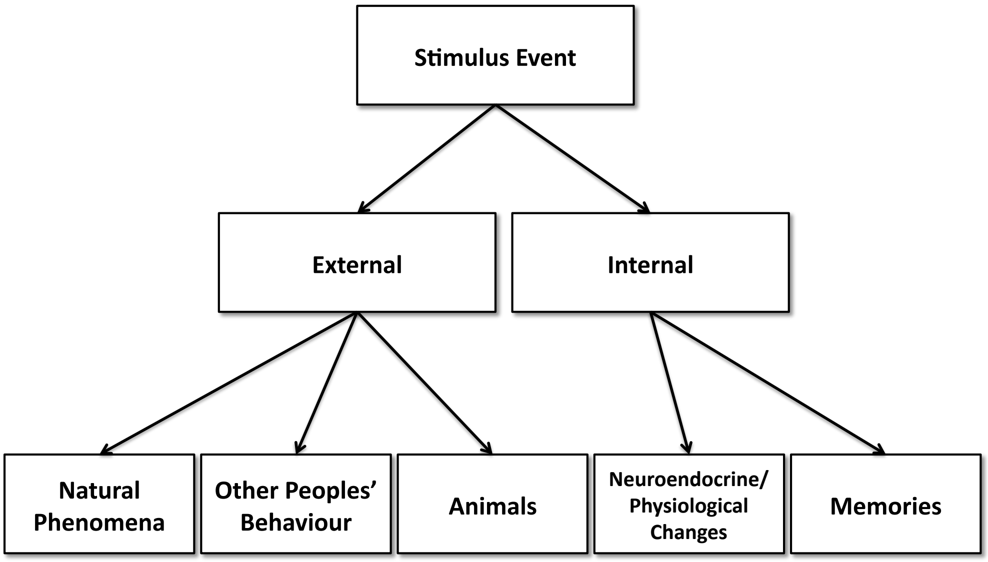

Figure 3.1: Sources of emotion stimuli

Tomkins (1984) views emotions as the primary motivation system that can amplify other physical and bodily functions, e.g., interference with breathing causes terror that leads to the struggle for air. Plutchik (1980) also stresses the evolutionary link of emotion with instinctive behaviour in animals, while Ekman (1999a) regards emotions as psychosomatic states that have evolved over time due to their adaptive value in dealing with prototypical life tasks. He suggests that emotions' primary function is to mobilise an organism to respond quickly to prototypical events, similar to those that were encountered in the past.

3.3 Anatomy of Emotion

The two dominant views on emotion structure are the discrete and continuous approaches. Discrete emotion theorists, following Darwin's work, suggest the existence of six or more basic emotions (happiness, sadness, anger, fear, disgust, and surprise), which are universally displayed and recognised (Darwin, 2005; Ekman, 1992; Ekman, 1999a). The arguments for the existence of basic emotions include cross-cultural universals for facial expressions and antecedent events, and presence of these emotions in other primates. Experiments in many countries, including countries isolated from media, show that people express and recognise basic emotions the same way (Ekman and Friesen, 1975).

There is no agreement on which emotions qualify as basic, but the list typically includes fear, anger, disgust, happiness, sadness, and surprise (Plutchik, 1980; Ekman, 1992). Other emotions are seen as combinations of these basic emotions or as socially learned variants of these emotions, e.g., grief, guilt and loneliness are all variants of basic sadness (Bower, 1992). In the context of MMHCI research, the theory of basic emotion implies that: i) emotional experiences can be measured on all stages of human interaction with systems, ii) accurate translations of emotional expressions and predictions based on these translations can be made, and iii) systems can incorporate characters depicting basic emotions that users can recognise accurately.

The continuous approach assumes the existence of two or more dimensions that describe and distinguish between different emotions (Russell, 1994; Russell and Mehrabian, 1977; Russell and Steiger, 1982; Barrett and Russell, 1999). Support for the dimensional emotion theories comes from physiological correlates with emotional stimuli, such as heart rate and skin conductance levels. The first dimensional model was developed by White (2004), who applied both introspective and experimental methods to study subjective experiences. The model is supported by other research, which revealed that people tend to perceive all kind of meaning in terms of valence and activation (Scherer, 2002). Russell (1994) proposed the use of independent bipolar dimensions of pleasure-displeasure, arousal, and dominance-submissiveness, rather than a small number of discrete emotion categories. Valence represents the pleasantness of the stimuli along a bipolar continuum, between a positive and a negative pole, while arousal indicates the intensity of the emotion. This dimensional taxonomy of emotions treats all emotion categories as varying quantitatively from one another (Russell and Steiger, 1982) and represents their relationships as distances within the affect space.

3.4 Methods for Measuring Emotions

Considering the plethora of definitions one cannot expect a single standard method for measuring emotions. Scherer (2005) argues that due to the component nature of the phenomenon only an assessment of all components involved can facilitate a comprehensive and accurate understanding of an emotional episode. This view suggests the observation of the following emotion-cognitive components: (i) changes in the appraisal processes (at all levels of the nervous system), (ii) responses produced in the neuroendocrine, autonomic, and somatic nervous system, (iii) motivational changes brought by the appraisal process, (iv) facial, vocal and bodily indications, and (v) nature of the subjectively experienced emotional state that relates to the above component changes. Though such accurate measurement of emotion has not been accomplished yet, significant progress has been made in measuring its individual components. In the next sections I will discuss observer (facial expression analysis, speech analysis, etc.) and self-report methods that can be applied in the study of emotion (Larsen and Fredrickson, 1999).

3.4.1 Physiological Signal Processing

One of the emotion components is physiological arousal, which corresponds to the physiological changes (e.g., respiratory and cardiovascular accelerations and decelerations, muscle spasms, etc.) that occur in emotion episodes and cannot be easily faked (Cornelius, 2000). Physiological methods involve monitoring body responses to emotional stimuli. The significance of these changes is due to the emotion-eliciting event causing a disturbance to the ongoing regulation and smooth behavioural coordination of the organism, and its preparation for appropriate response (e.g., fight or flight).

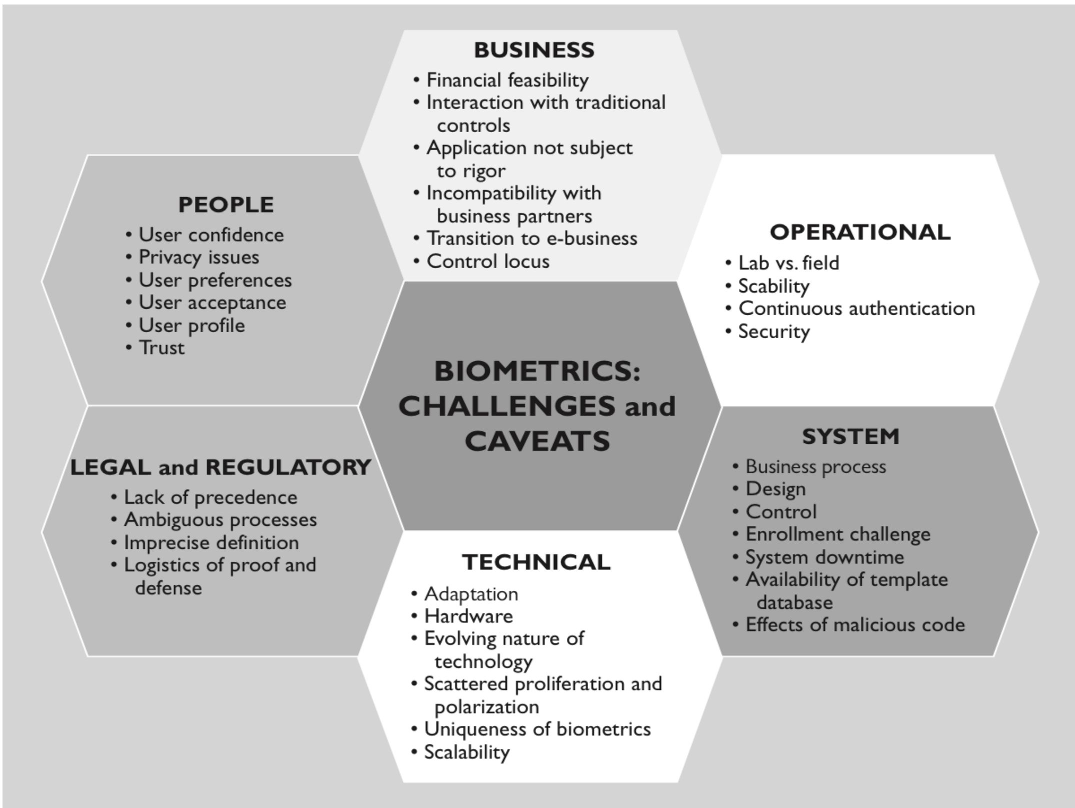

Figure 3.2: Challenges of biometric technologies (Chandra and Calderon, 2005)

With the emergence of the ubiquitous computing era (Weiser, 1993; Weiser, 1994), biometric technologies received increased attention and their application, especially in areas such as surveillance and authentication, became more broad. Physiological measurements by definition identity an individual uniquely, based on some features of their biological makeup (Bullington, 2005). These features can be behavioural (such as the way somebody writes their signature) or physiological (fingerprints, the patterning of the iris, etc.). Some of the most popular physiological features are: (i) voiceprint, (ii) hand geometry, (iii) fingerprint, (iv) iris, (v) retinal, (vi) galvanic skin response (GSR), (vii) electrocardiogram (ECG), (viii) electroencephalogram (EEG), (ix) skin temperature, (x) heat flux, and (xi) blood volume pulse (BVP).

Researchers can detect emotional arousal by scanning brain activity, pulse rate, blood pressure or skin conductance. For example, changes in the electrical properties of the skin, due to the activity of the sweat glands, is physically interpreted as conductance (Boucsein, 1992). The sweat glands, which are distributed all over the skin, receive input from the sympathetic nervous system, making it a good indicator of the level of emotional arousal due to external sensory or cognitive stimuli. Skin temperature, which reflects the autonomic nervous system activity, is also another reliable indicator of the underlying affect state.Variations in the skin temperature are mainly the result of localised changes in the blood flow, caused by vascular resistance or arterial blood pressure (Kataoka, Kano, Yoshida, Saijo, Yasuda and Osumi, 1998). These sensory channels have been used to measure negative emotions, such as stress and anxiety (Wilson and Sasse, 2000), and user interest towards multimedia content (Soleymani et al., 2008; Smeaton and Rothwell, 2009) and games (Chanel, Rebetez, Bétrancourt and Pun, 2008). The procedures for collecting physiological measures vary between a simple sensor on a finger (for monitoring pulse rate and skin conductance) to more invasive sensors, such as ECG or EEG.

Biometric systems must contend with a variety of problems, when applied in real-world applications (Figure 3.2). According to Jain and Ross (2004), physiological signals are often noisy. A fingerprint with a scar or a voice that has been altered due to a sore throat are examples of noisy data. Moreover, noisy data can also result from improperly maintained or defective sensors, bad environmental conditions, and other. In addition, there are intra-class variations, which refer to the difference between the biometric data acquired from an individual during authentication and that used to generate the template during enrolment. A third problem is diversity. The physiological responses among individuals is expected to be more diverse and, therefore, it is generally harder to determine whether these transitions occur due to a change in the affect state or other factors, e.g., cognitive processes, sensory stimuli, etc., However, there may also be large similarities in the feature sets used to represent these features. Lastly, there is the issue of non-universality. Although every individual is expected to possess certain biometric traits, in reality it is possible for a subset of the user population not to possess a particular biometric. For example, a fingerprint biometric system may be unable to extract features from the fingerprints of an individual whose quality of the ridges is extremely poor.

Recently, biometrics became increasingly popular, especially in the areas of affective computing and MMHCI. Heishman, Duric and Wechsler (2004) applied eye region biometrics in single participants to determine their affective and cognitive states in HCI scenarios. Moore and Dua (2004) provided findings of the application of biometrics (galvanic skin response) for the effective interaction and communication with a biometric-based control interface. Bullington (2005) explored the intersection of several groups of technologies (surveillance camera networks, ubiquitous computing, biometrics, face recognition, and affective computing) and introduced the deployment of affect recognition in three different scenarios. Snelick, Indovina, Yen and Mink (2003) examined the application of multi-modal biometric systems in large scale populations and introduced a methodology for testing the performance of such systems. Finally, Swindells, MacLean, Booth and Meitner (2006) presented results from two user studies, where self-reports and biometric measurement techniques were compared and contrasted.

3.4.2 Facial Expression Analysis

Research indicates that emotions are primarily communicated through facial expressions, rather than bodily gestures (Ekman and Friesen, 1975). Facial cues (smiles, chuckles, smirks, frowns, etc.) are an important aspect of social interactions (Russell et al., 2003); they help determine the focus of attention (Pantic and Rothkrantz, 2000b) and regulate human interactions with the surrounding environment (Figure 3.3). The research performed in the area of emotion recognition through facial expression analysis has been largely inspired by Darwin's pioneering work (Darwin, 2005). Another proponent of emotion theory is Ekman, who's work involved the development of a classification system of discrete emotion categories (Ekman, 1999a; Ekman, 1999b; Rosenberg and Ekman, 1993), as well as formulation of the Facial Action Coding System (FACS). FACS is an objective and unobtrusive technique for measuring muscular movement in terms of distances between feature points on the face (known also as action units), and was inspired by the earlier work of Hjortsjö (Russell et al., 2003). Ekman's research on emotions laid the foundations for the development of automatic extraction and analysis of affective information, conveyed through facial expressions.

Facial expressions are the result of facial muscle contractions, which cause temporary deformations of facial skin and certain landmark features (eyebrows, nose, mouth, etc.). FACS is based on recognising facial expressions of six universally distinguished emotions:

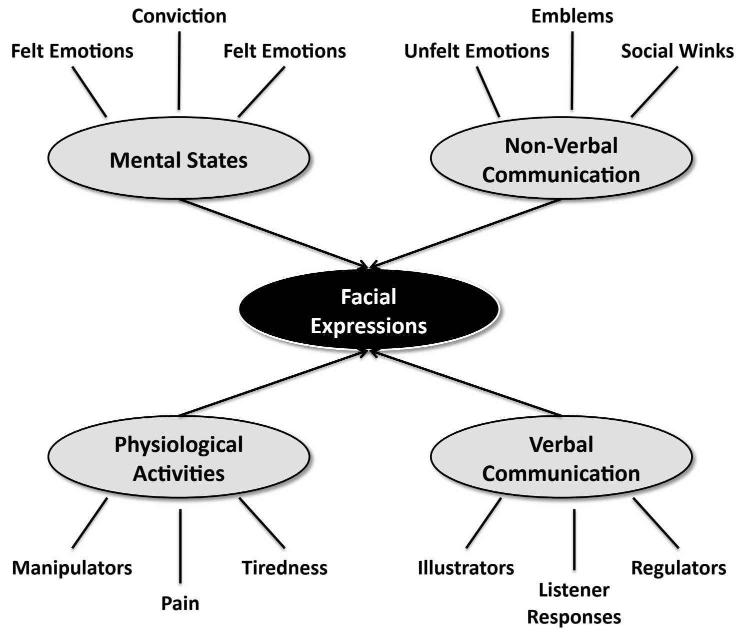

Figure 3.3: Sources of facial expressions (Fasel and Luettin, 2003)

fear, surprise, sadness, happiness, anger, disgust, and their combinations. The intensity of the emotion can be determined indirectly by the presence and degree of changes in all facial regions associated with it. For example, sadness is usually expressed through the brow, eye and mouth areas. In sadness, the inner corners of brows are drawn up, skin below the eyebrow is triangulated with the inner corner up, upper eyelid inner corner is raised, corners of lips are down or the lip is trembling (Ekman, 2003a). Among the benefits of the FACS method are: i) high reading accuracy rates, ii) use of non-obtrusive and common laboratory equipment, such as video camera, and iii) high validity that is confirmed by correlations with physiological measures (e.g., increased heart rate that coincides with surprise and disgust). Another benefit of FACS is the fact that it can be programmed into computer systems to automatically recognise people's emotions (Picard, 1997; Cohn and Kanade, 2006).

There are two major approaches that are informed by FACS and are based on automatic facial feature extraction and representation: (i) the feature-based approach, and (ii) the region-based approach. Systems that apply the first approach analyse appearance-based changes in localised features of the face, such as the corners of the mouth, eyebrows, etc. (Sebe, Lew, Cohen, Sun, Gevers and Huang, 2004; Valenti, Sebe and Gevers, 2007; Pantic and Patras, 2006; Wang, Ai, Wu and Huang, 2004). In the second approach the feature vectors are obtained from segmented regions around each fiducial point of the face, such as eye/mouth regions (Essa, 1995; Essa and Pentland, 1997).

Furthermore, we can distinguish two types of classification schemes (Jaimes and Sebe, 2007): static and dynamic. Static classifiers process each image individually to one of the facial expression categories, while dynamic classifiers process an image sequence (or a set of video frames) and apply classification by analysing the temporal patterns of the extracted regions or features. Static classifiers do not require extensive knowledge of the object of analysis, and tend to be generally fast and simple. However, they are considered more error-prone and unreliable when dealing with many different views of the same subject (Fasel and Luettin, 2003). Dynamic classifiers, on the other hand, are better at detecting changes in facial expressions of different individuals and, therefore, are more suitable for individual-independent studies.

3.4.3 Speech Analysis

The empirical investigations of the effects of emotions on speech begun in the early 20th century, when the first systematic attempts to detect emotional arousal in spoken statements were made (Mendoza and Carballo, 1999). Since then scientific interest in the vocal attributes of moods, affective (Scherer, 2003) and cognitive states has increased. In recent years research on vocal expression patterns of the naturally occurring emotions has produced a substantial number of theories and models of speech communication. Brunswik's functional lens model of perception (Figure 3.4) is an example of such a model (Mitroff, 1974). The model suggests that vocal communication of emotion is initiated with an encoding, or expression, of an emotional state by certain voice and speech characteristics that are susceptible to objective signal measurement. The assumption made here is that that there are some acoustic correlates of emotion in the acoustic parameters (e.g., respiration, phonation) that can provide insightful cues about the speaker's affective state.

According to Pantic and Rothkrantz (2003), the auditory features that are most often used in speech analysis are: (i) pitch, (ii) intensity, (iii) speech rate, (iv) pitch contour, and (v) phonetic features. Pitch corresponds to the rate at which vocal cords vibrate and determines the frequency of the acoustic signal, while intensity refers to vocal energy. Variations in voice pitch and intensity usually have a linguistic function, such as over-stressing or under-stressing certain words (Cowie, Douglas-Cowie, Tsapatsoulis, Votsis, Kollias, Fellenz and Taylor, 2001). When, for example, a person is experiencing anger, fear or joy, the sympathetic nervous system becomes aroused, resulting in a heart rate and blood pressure increase that produces mouth dryness and occasional muscle tremors. Speech is then

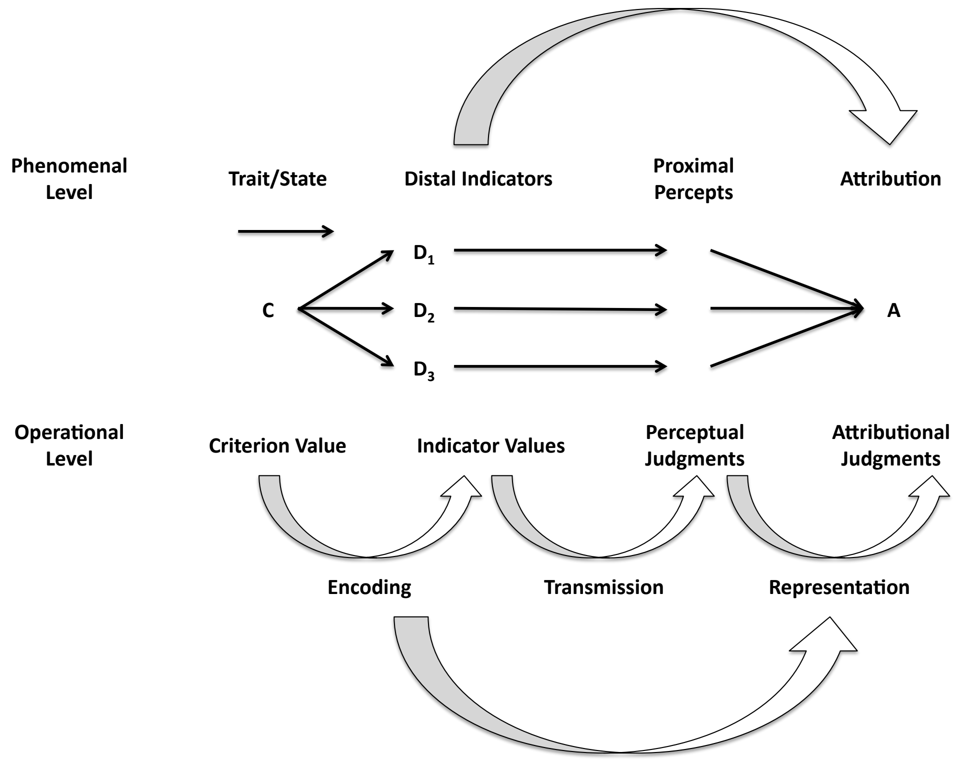

Figure 3.4: A Brunswikian lens model of the vocal communication of emotion (Mitroff, 1974)

characterised by loudness, increased speech rate and strong, high frequency energy (Breazeal, 2001). Speech rate represents the number of spoken words within a time interval. Finally, pitch contour corresponds to pitch variations described in terms of geometric patterns, and phonetic features of all types of sounds involved in a speech (e.g., vowels, consonants and their pronunciation). Table 3.1 illustrates correlations between emotion and acoustic parameters from the Murray and Arnott (1993) review.

According to Jaimes and Sebe (2007), existing vocal affect analysers have four limitations: (i) singular classification of input audio signals into a few discrete emotion categories, (ii) context-independent analysis of the input audio signal, (iii) analysis of vocal expression information only on short time scales (thus inferences about moods and attitudes are almost impossible to obtain), (iv) assumptions about the quality of the test data (noise-free recordings, short sentences with intermediary pauses, clear speech, etc.).

Systems that analyse speech, along with other representations of affect, are described by Go, Kwak, Lee and Chun (2003), Schuller, Lang and Rigoll (2002), Sebe, Bakker, Cohen,

	Anger	**Happiness**	**Sadness**	**Fear**	**Disgust**
Rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much faster
Pitch Average	Very much higher	Much higher	Slightly lower	Very much higher	Very much lower
Pitch Range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy, chest	Breathy, blaring tone	Resonant	Irregular voicing	Grumble chest tone
Pitch Changes	Abrupt on stressed	Smooth, upward inflections	Downward inflections	Normal	Wide, downward terminal inflects
Articulation	Tense	Normal	Slurring	Precise	Normal

Table 3.1: Emotions and Speech Parameters (Murray and Arnott, 1993)

Gevers and Huang (2005), Corradini, Mehta, Bernsen and Martin (2003), Schapira and Sharma (2001), Yoshitomi, Sung-Ill, Kawano and Kilazoe (2000), Song, Bu, Chen and Li (2004), Schuller, Arsic, Wallhoff and Rigoll (2006), and Zeng, Hu, Fu, Huang, Roisman and Wen (2006). These systems achieve an accuracy rate of 72% - 85% when detecting one or more basic emotions from noise-free audiovisual input (Jaimes and Sebe, 2007). These accuracy rates outperform the equivalent scores of human emotion recognition, which range between 55% - 70% in neutral content speech (Pantic and Rothkrantz, 2003).

3.4.4 Gesture Recognition

Whether body movements or gestures are indicative of specific emotions is a subject under debate. Some studies suggest that gestures are indicative of the intensity of emotion, but not its type. Other studies (Boone and Cunningham, 1998; Meijer, 2005; Wallbott, 1998) provide evidence that associate certain body movements to specific emotions. This approach follows Darwin's view of humans' genetic predisposition to exhibit certain patterns of bodily movements during the expression of emotions (Darwin, 2005). Body movements, and specifically hand gestures (Chen, Fu and Huang, 2003; Castellano, Villalba and Camurri, 2007; Caridakis, Castellano, Kessous, Raouzaiou, Malatesta, Asteriadis and Karpouzis, 2007), have recently attracted the attention of the HCI community (Ambady and Rosenthal, 1992). Gunes and Piccardi (2007) fused facial expressions and body gestures for bimodal emotion recognition. In their study they provided a list of expressive gestures and their correlation to the emotion categories. A list of emotions recognised by the changes that occur on the body (Gunes and Piccardi, 2007) are presented in

Table 3.2. Among all forms of gestures, hand gestures have been one of the most common and natural communication media (Chen et al., 2003) and a critical link between our conceptualising capacity and our linguistic abilities. Hand gestures are used for a variety of reasons, such as to point at an object, indicate a feeling (Kapoor, Picard and Ivanov, 2004; Balomenos, Raouzaiou, Ioannou, Drosopoulos, Karpouzis and Kollias, 2005), or communicate a stance.

Anxiety	Hands close to the table surface; fingers moving; fingers tapping on the table
Anger	Body extended; hands on the waist; hands made into fists and kept low, close to the table surface
Disgust	Body backing; left/right hand touching the neck or face
Fear	Body contracted; body backing; hands high up, trying to cover bodily parts
Happiness	Body extended; hands kept high; hands made into fists and kept high
Uncertainty	Shoulder shrug; palms up

Table 3.2: List of bodily emotions and the accompanying changes that occur on the body when they are displayed

Most of the recent work in hand gesture recognition can be grouped into: (i) glove-based, and (ii) vision-based (Chen et al., 2003). Glove-based gesture recognition requires the user to wear an unobtrusive hand glove device, which communicates gestures to a computer system through a set of wires. This approach is based on the 3-D spatial description of hands. Vision-based gesture recognition method relies on appearance of hands in images and applies appearance-based techniques, while glove-based recognition applies model-based techniques for gesture analysis. McAllister, McKenna and Ricketts (2002a) outlined the major difficulties in using gesture recognition techniques, which are related to the following factors: i) the hand's jointed physical structure often results in self-occlusion, which makes it harder to model, ii) many gesture recognition applications cannot track hands under poorly controlled and varying lighting conditions, iii) tracking both hands at the same time demands a solution to the temporal matching (data association) problem, and a method for dealing with the temporary occlusion of one hand by the other. Finally, gesture recognition can be hindered by clothing or other objects. Examples of the studies that used gesture recognition include McAllister, McKenna and Ricketts (2002b), Ng and Ranganath (2002), Brewster, Lumsden, Bell, Hall and Tasker (2003), Chen et al. (2003), Camurri, Lagerlöf and Volpe (2003), Kapoor et al. (2004), Kaliouby and Robinson (2004), Kirishima, Sato and Chihara (2005), Licsár and Szirányi (2005), Isbister, Höök, Sharp and Laaksolahti (2006), Morency and Darrell (2006), Gunes and Piccardi (2007), Morency, Sidner, Lee and Darrell (2007), Tsalakanidou, Malassiotis and Strintzis (2007).

3.4.5 Eye Tracking

Gaze is the direction to which the eyes are pointing in space. A body of work in psychology and neuroscience has shown that the gaze direction and eye fixation are strong indicators of attention (Duchowski, 2002; Salojärvi, Kojo, Jaana and Kaski, 2003; Salojärvi, Puolamäki, Simola, Kovanen, Kojo and Kaski, 2005). The neuroanatomical basis for the importance of gaze direction lies in the structure of the retina and the visual pathway (accurate viewing is possible only in the central fovea area). Eye tracking technology is considered a valuable component of interactive systems, since it can provide a quantitative measure of time overt attention and help infer higher order cognitive processes from eye movements.

Eye tracking systems can be categorised into: (i) wearable, (ii) non-wearable, (iii) infrared-based, and (iv) appearance-based. Infrared-based tracking systems project a collimated beam of infrared light at the front surface of the eyeball, producing a bright glint or corneal reflection. The corneal reflection and outline of the pupil can then be used to compute the orientation of the eye and predict the direction of sight (Jacob, 1995; Zhu, Fujimura and Ji, 2002). Appearance-based systems use a camera that observes the eyeball and determine the direction and orientation of the eyesight by applying image processing algorithms (Ware and Mikaelian, 1987; Hansen, Hansen, Niels and Stegmann, 2002; Tan, Kriegman and Ahuja, 2002). Due to the computational cost of processing two streams simultaneously, a single high resolution image of one eye has been proposed (Wang, Sung and Venkateswarlu, 2003). Overall, infrared-based systems are regarded as more accurate then appearance-based, yet there are concerns over the safety of prolonged eye exposure to infrared light.

Most wearable eye trackers have been developed for desktop application (Dickie, Hart, Vertegaal and Eiser, 2006). However, due to the advances in hardware development, researchers have been able to apply wearable eye tracking in novel ways (Smith, Vertegaal and Sohn, 2005). Vertegaal, Dickie, Sohn and Flickner (2002) integrated eye tracking in a prototype attentive cell phone, which uses a low-cost EyeContact sensor and speech analysis to detect whether its user is in a face-to-face conversation. Rothkopf and Pelz (2004) applied eye tracking in natural tasks, where participants were able to move freely

Modality	**Sensor**	**Is it socially communicated**	**Comments**
Facial Activity	Video	Yes	Facial expressions can differ significantly from genuinely felt feelings
	Infrared Video	Highlights pupils & specularities	Usually works better than ordinary video when head moves (better eye detection)
	Thermal Video	No	Being explored to detect stress and other changes related to deception and frustration
Posture Activity	Force-sensitive resistors	Yes, but not as pressure	Good results discriminating level of interest in students in computer learning interactions
Hand Tension & Activity	Force-sensitive resistors or Sentograph	Varies; depends on gesture	Can be sensed from handling of mouse, steering wheel, etc., and pressure has been shown to be higher during a frustrating task
Gestural Activity	Electromyogram electrodes	Visibility varies	Shown for expression sensing in conducting music; other gestures largely unexplored w.r.t. expression
Vocal Expression	Microphone	Yes	Most methods are great for discriminating arousal but not for valence; limited to times when user is speaking
Language and choice of words	Text analysis tools	Yes	Can be used with interfaces requiring textual input; promising for valence; trivial to sense scripted dialogue moves
Electrodermal Activity	Electrodes (can also be clothing snaps, metallic fabric, etc.)	No; except perhaps sweaty palms	Good at detecting changes in arousal but doesn't distinguish positive/negative, and can also be triggered by non-affective changes

Table 3.3: Body-based measures of affect (partial set of examples) (Picard and Daily, 2005)

in the environment. The proposed device was able to record a video of the surroundings of the participant without being limited to a small field of view, as obtained by most cameras employed in eye tracking. Furthermore, the device could recover the head movements, as well as offer additional information about the type of eye movement and the gaze change in the environment coordinates. Finally, Dickie, Vertegaal, Fono, Sohn, Chen, Cheng, Shell and Aoudeh (2004) applied eye tracking to record video for blogging, and Pelz, Canosa, Kucharczyk, Babcock, Silver and Konno (2000) used a wearable eye tracker to explore the role of eye movements with respect to visual perception and hand-eye coordination in natural tasks. While wearable systems are the most accurate, they also appear to be the most intrusive. In addition, most non-wearable systems require calibration for each individual.

Several key problems and solutions have been identified in interactive eye tracking systems, such as the Midas Touch problem (Dickie et al., 2006). For users with disabilities, eye tracking can be an indispensable form of communication. However, its application in more general interactive settings raises a question as to whether it is necessary to overload a perceptual organ, such as the eye, with a motor task (e.g., mouse-like pointing). As a complementary interface modality, eye tracking may potentially serve as a predictor of the user's cognitive states in affective interfaces. Even though eye-tracking related features are not directly associated to the affective state of the user, they remain good indicators of attention, that can help determine potential sources of emotional stimuli and, furthermore, identify their importance to the observer.

3.4.6 Self-report Methods

While physiological response patterns and expressive behaviour can be easily observed and analysed to determine the underlying emotions there are no objective methods of measuring the subjective experience. In many cases researchers simply ask participants to describe the nature of their experience through the use of self-report techniques. The self-report methods rely on assumptions that individuals are able and willing to recognise and report their emotions. The reliability and validity of the measures are evident from the high correlations of the reports to the quality of the physical stimuli and neurological activities of the brain (Kahneman, 2000). Momentary reports are considered the most accurate; however, there are techniques for improving the accuracy of retrospective reports. While they may be subject to participant's bias, self-report methods are efficient and easy techniques for measuring emotions.

The two major methods of self-reports are: (i) the discrete emotions approach, and (ii) the dimensional approach. The first approach relies on the categorisation that is reflected in the classification of the semantics for emotions in natural language. The fact that language-based categories appear to correspond to unique emotion patterns provides sufficient justification for accepting the latter structure. In most cases, a list of emotion terms is provided to a respondent who must determine which term describes better his/her emotional experience, rate the intensity of emotion and, finally, state how much of that emotion has been experienced.

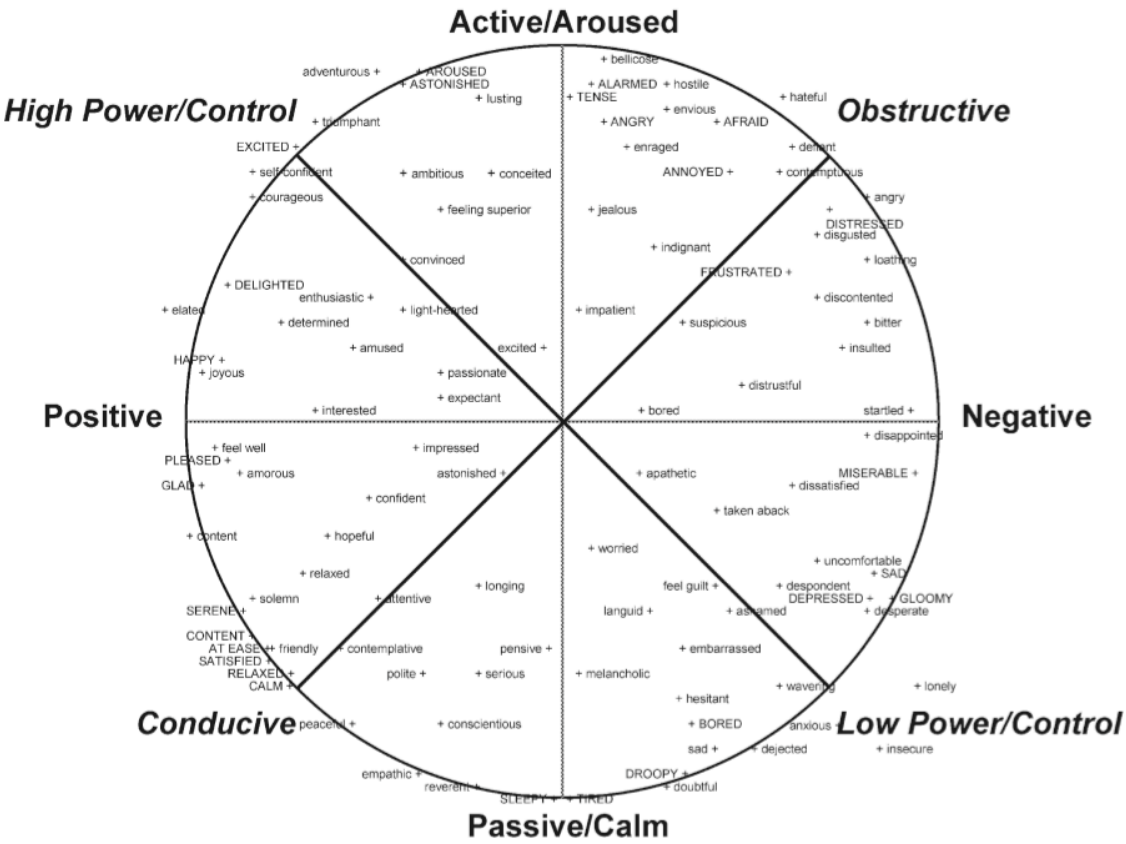

Figure 3.5: Alternative dimensional structures for the semantic space for emotions (Scherer, 2005)

While ensuring efficiency and standardisation of data collection, discrete emotion self-reports have several disadvantages, including: i) the possibility that one or several response alternatives may bias the respondent to choose them, ii) the situation when a respondent wishes to refer to a category that is not provided on the list, or iii) the situation when a respondent may be unfamiliar with the labels chosen by a researcher (Scherer, 2005). Russell and Steiger (1982) argue that, when using natural language, people classify emotions by means of a taxonomy, but cannot explicitly describe it. Thus, the taxonomy is implicit in the same sense that the syntactic rules of language are implicit. Lastly, while the information obtained from this approach appears to be intuitive and easily interpretable, there are issues of comparability of results between studies that employed different lists of emotion labels (Scherer, 2005).

The second approach was established by Wilhelm Wundt (Scherer, 2005; Sander, Grandjean and Scherer, 2005) and it is based on the use of latent dimensions. Wundt suggests the use of a three-dimensional space, formed by the dimensions of valence (positive-negative), arousal (calm-excited), and tension (tense-relaxed). Given this approach, a respondent can report his/her emotional experience by simply indicating coordinates on a three-dimensional space. Due to the difficulty of consistently identifying a third dimension from arousal or excitation, researchers often apply only two out of the three dimensions, thus forming a two-dimensional surface (arousal-valence space).

This approach is quite straightforward, simplistic and, above all, provides interval data that can be readily used in statistical processing (Russell and Steiger, 1982). However, the results lack the intuitiveness of the discrete emotions approach and are limited to degrees of positive or negative valence or arousal (Figure 3.5). Another shortcoming of the dimensional approach is the inherited ambiguity of valence and arousal, i.e., it is not always clear whether a valence judgement is indicative of the appraisal of the nature of the stimulus or the feeling induced by it. Most importantly it is very difficult, if not impossible, to distinguish the intensity of an emotion from bodily excitation. As a result, extremely intensive happiness may be characterised by high arousal, while intense sadness may be accompanied by very low arousal.

To mitigate the disadvantages of self-report methods, researchers often choose to use free-response reports that allow participants to identify their emotions using words or expressions that best represent their experiences. This technique provides a high level of specificity, which can be useful in studies where accuracy and explicitness are considered important. However, it is difficult to analyse such free-response data in a quantitative, statistical manner (Scherer, 2005).

3.5 Emotions in Information Retrieval Research

Until recently the IR community excluded emotion research from its agenda. However, during the recent years a growing number of studies have investigated the role of various emotive variables in the context of IR, revealing an increased interest in emotion-related IR research.

3.5.1 User-based Studies

A number of studies have investigated the affective dimensions of online information seeking behaviour by applying a range of user-centred techniques, such as facial expression analysis or physiological signal processing. Mooney, Scully, Jones and Smeaton (2006), examined the role of searchers' emotional states in an attempt to improve data indexing for and within the search process. Users' physiological responses to emotional stimuli were recorded using a range of metrics (galvanic skin response, skin temperature, etc.). The study provided evidence in favour of using physiological signal processing for studying searchers' emotions.

Kim (2008) studied the relationships between search tasks, user emotion control and performance during the web search. The experiment involved completion of different search tasks of varying scope (specific task vs. general task), and reporting users' expectations of problem solving. The results indicated that both tasks and emotion control impact user search behaviour. The author discusses ways of releasing cognitive and affective burden on the searcher by offering information literacy education and improving interface design.

An investigation into the role of emotions in the information seeking process by Arapakis et al. (2008) has provided initial evidence of the effect of search task difficulty on users' emotions. More specifically, the findings indicated a progressive transition from positive to negative valence as the degree of task difficulty increased. The study applied automatic facial expression analysis to infer users' affective state throughout the tasks.

Facial expression analysis was used in a study of digital libraries' search by Lopatovska and Cool (2008). The authors found that, during the search, the participants' faces expressed primarily dislike and variations of this emotion, while in one participants most of the positive emotions corresponded with the time when an assistant entered the room. The study also found a wide variation in individual levels of emotional expressivity (e.g., one participant's face expressed 57 intense emotions, while other participant's face expressed only 9 emotions during the same period of search time).

Arapakis et al. (2009) developed a novel video search environment that applied real-time facial expression analysis to aggregate information on users' affective behaviour. The collected information was used to classify the topical relevance of the perused videos and enrich the user profiles. The value of the system lies in the combination of different modalities (facial expression data, interaction data, etc.), the integration of affective features into the employed profiling techniques, and, finally, the facilitation of meaningful recommendations of unseen videos.

Finally, Lopatovska (2009) used the Positive Affect and Negative Affect Scale (PANAS) to measure searchers' affect between search tasks. The PANAS (Watson, Clark and Tellegen, 1988) comprises of two 10-item scales that measure positive affect (extent to which a person feels enthusiastic, active, alert, etc.) and negative affect (extent to which a person experiences subjective distress, including anger, contempt, disgust, guilt, fear, nervousness, etc.). The study showed that affect did not change significantly during the course of the search and was not significantly influenced by the search process. A study that investigated subjective variables of the information search process found that better mood before the search and during the search correlates with better mood after the search, but also correlates with the worse search outcomes and lower satisfaction. The latter finding perhaps suggests that it pays off to feel some "pain" during the search in order to "gain" quality outcomes (Gwizdka and Lopatovska, 2009).

3.5.2 Content-based Studies

Recent studies have also examined the affective dimensions of information seeking by focusing on content features, rather than user behaviour. A content-centred approach was adopted by Chan and Jones (2005) and Hanjalic and Xu (2005). Chan and Jones (2005) extracted a range of affective features from multimedia audio content and annotated it using a set of labels with predetermined affective semantics. The audio features, which consisted of speech, music, special effects and silence, were analysed in terms of the affective dimensions of arousal and valence. These measurements were combined to form a plot, known as the affect curve. Similarly, Hanjalic and Xu (2005) modelled video content using a selection of low level audio (signal energy, speech rate, inflection, rhythm duration, voice quality) and visual features (motion). This framework was based on the dimensional approach to emotion, with the video content represented as a set of points in a two-dimensional (arousal and valence) affect space that reliably depicts expected emotional transitions across the video (as perceived by a viewer).

Soleymani et al. (2008) proposed an approach to affective ranking of movie scenes which is based on viewers' affective responses, as well as content based multimedia features. The latter features captured important aspects of the events that characterise every scene and were correlated with the viewers' self-assessments of arousal and valence. Furthermore, the provided evidence suggests that peripheral physiological signals can be used to characterise and rank video content, even though the variation of users' affective responses emphasises the need for more personalised emotional profiles.

Finally, a number of IR studies are focusing on emotions as descriptors of information and are forming a new area of emotional IR research. Certain information objects, such as music and images, induce emotions that help to describe and ultimately retrieve these objects. Lee and Neal (2007) developed a system for tagging emotions and their intensity induced by music. Schmidt and Stock (2009) followed a similar approach for tagging images, using basic emotion tags such as anger, disgust, fear, happiness and sadness. The study findings suggest that collective emotion tagging can provide a new way for retrieving relevant information.

3.6 Experimental Methodology

In the previous sections I reviewed literature on the theories of emotions, methods for studying emotions, as well as examples of user-centred and content-based studies of emotions from the fields of IR and LIS. Next, I provide an overview of the experimental methodology (Sections §3.6.1) and the modalities used for data collection (Sections §3.6.2 and §3.6.3), discussed in Chapters 4, 5, 6, 7 and 8.

3.6.1 Experimental Conditions

By definition an experimental study introduces the participants to an artificial situation that takes place in a laboratory setting, therefore lacking the ecological validity of a naturalistic study. In addition, when applying a multi-modal analysis several critical issues arise. According to Sebe, Lew, Sun, Cohen, Gevers and Huang (2007), emotional expressions are highly idiosyncratic in nature and may vary significantly from one individual to another (depending on personal, familial or cultural traits). Furthermore, spontaneous expressive behaviour may not be easily elicited, especially when participants are aware of being recorded. Most studies suffer from a significant trade-off between the participants being aware they are monitored (open recording) and not possessing that knowledge (hidden recording), therefore acting more spontaneously and naturally. 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 I devised a user study, which mitigated most of the unwanted effects. In my approach I : (i) employed a facial expression recognition system of reasonably robust performance and accuracy across all individuals, (ii) whenever possible, I applied hidden recording to increase the chance of observing spontaneous behaviour, and (iii) made the experimenters' presence in the laboratory setting less noticeable. The primary goal was to create sufficient ground truth where facial expressions, among other sensory input, would correspond to the underlying emotional state of every participant.

3.6.2 Facial Expression Analysis

Facial expressions have been associated in the past with universally distinguished emotions, such as happiness, sadness, anger, fear, disgust, and surprise (Ekman, 2003a). Recent findings also indicate that emotions are primarily communicated through facial expressions (Pantic and Rothkrantz, 2000a) and are generally regarded as inherent elements of human social interaction. Affective information conveyed through the visual channel can be crucial to human judgement and offer valuable insights to the observer. The face provides conversational signals (smiles, chuckles, frowns, etc.) that do not only clarify our current focus of attention (Pantic and Rothkrantz, 2003) but also regulate our interactions with the surrounding environment and the organisms that inhabit it.

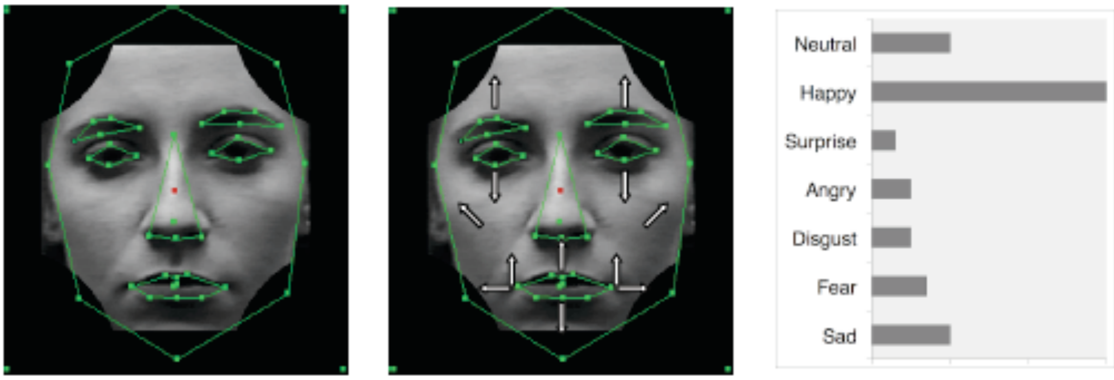

Figure 3.6: eMotion

The user studies presented in the following chapters applied automatic facial expression analysis using the feature-based system described in Valenti et al. (2007). Automatic systems are an alternative approach to facial expression analysis (Pantic, Sebe, F. and Huang, 2005) and have exhibited performance comparable to that of trained human recognition (87%). The facial expression detection takes place as follows: initially, eMotion locates certain facial landmark features (such as eyebrows, the corners of the mouth, etc.) and constructs a 3-dimensional wireframe model of the face that consists of a number of surface patches wrapped around it (Figure 3.6). After the construction of the model, head motion or any other facial deformation can be tracked and measured in terms of motion-units (MU's). The intensity and category of an emotion can then be determined indirectly by the presence and degree of changes in all facial regions associated with it.

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 Valenti et al. (2007) support the performance-related assumptions, although I refrain from claiming that such characteristics are of no importance, especially on an interpretation level.

Furthermore, eMotion applies a static classification scheme, which entails the processing of each frame independently from it's neighbouring frames and classifies it to one of the facial expression categories. Static classification is considered more error-prone and unreliable (Fasel and Luettin, 2003). However, it does not require an extensive knowledge of the object of analysis and is generally faster and simpler to implement. Several other issues were also taken into consideration, such as occlusion, illumination conditions, and other, which were potential noise factors.

Finally, automatic facial expression recognition does not account for context and, therefore, cannot perform a context-dependent interpretation of the data (Jaimes and Sebe, 2007). Fasel and Luettin (2003) 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 behavioural cues are far more accurate than those that are based on the body or the tone of the voice alone (Pantic and Rothkrantz, 2003). This suggests that affective information conveyed through the visual channel is crucial to human judgment and offers valuable insights to the observer. 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 automatic facial expression analysis have been used only as cues for emotion recognition and not as the ground truth itself. For a more detailed presentation of the above issues the reader is referred to (Pantic and Rothkrantz, 2003; Jaimes and Sebe, 2007).

3.6.3 Physiological Signal Processing

Emotions can be expressed through several sensory channels and are reflected by a series of more or less observable cues, such as localised changes in the electrodermal activity, variations in the skin temperature, and many more. It has been shown that transitions between emotional states are correlated with temporal changes in physiological states, which cannot be easily faked (Cornelius, 2000). In this work the participants' affective responses were measured using a range of physiological metrics, such as heart-rate, galvanic skin response and skin temperature. These modalities have been used in the past to measure negative emotions, such as stress and anxiety (Wilson and Sasse, 2000), and participants' interest towards multimedia content (Soleymani et al., 2008; Smeaton and Rothwell, 2009) and games (Chanel et al., 2008).

To measure participants' physiological signals I used Polar RS800 and BodyMedia SenseWear ® Pro_3 Armband. Polar RS800 Heart Rate Monitor consists of Polar RS800 Running Computer, a wrist-watch that displays and records the heart rate data, an elastic strap with two electrodes and Polar WearLink® 31, a wireless transmitter. The elastic strap is worn around the chest (bellow the chest muscles), allowing the built-in soft textile electrodes to detect the heartbeat and then transmit the heart rate signal to the running computer via the Polar WearLink® 31, which is attached to the strap.

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

3.7 Summary

Recent scientific findings suggest that emotion is a ubiquitous element of any human computer interaction (Brave, Nass and Hutchinson, 2005) and should be considered when designing usable and intelligent systems (Karat, 2003). Emotions not only regulate our social encounters but also influence our cognition, perception and decision-making. This chapter reviewed theories that define emotions and outline their role and importance in making computer-related tasks more pleasurable and natural for humans to understand.

I reviewed the two major categories of classical emotion theories: (i) cognitive, which stresses the importance of cognitive evaluation (appraisal) in establishing the meaning of stimuli and ways copying with it, and (ii) somatic, which emphasises somatic factors and describe expressions and perceptions of emotional experiences in terms of bodily responses. I, furthermore, discussed the two dominant views on emotion structure, namely: the discrete and dimensional approaches. The former supports the existence of six or more basic emotion categories that are universally displayed and recognised, while the latter suggests the representation of emotions in terms of a multi-dimensional space of arousal-valence.

Examples of studies that provide evidence for the importance of the role of emotions in human-computer interaction, and, in particular, search behaviour were also presented. Most IR studies of emotion take place in experimental settings and involve the use of a variety of methods and modalities. The emerging affective paradigm needs to become broader and richer, and to encourage research that reaches across the boundaries of the narrowly defined fields.

Finally, I have provided an overview of the experimental methodology and the modalities that were used for the collection of affective data, in the experiments presented next.

Chapter 4

User Study 1: An Investigation into the Role of Emotions in the Information Seeking Process

---

Preamble

User feedback is considered to be a critical element in the information seeking process, especially in relation to relevance assessments. Current feedback techniques determine 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 are critical aspects of cognition and decision-making. This chapter 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.

4.1 Introduction

User feedback is considered to be a critical element in the information seeking process (Harman, 1992). A key feature of the feedback cycle is relevance assessments, which have progressively become a popular practice in web searching activities and interactive IR. The value of relevance assessments lies in the disambiguation of user's information need and is achieved by applying various feedback techniques. Such techniques, as discussed in Sections §2.4.1 and §2.4.2, vary from explicit to implicit and help determine the topical relevance of the retrieved items.

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 (Koenemann and Belkin, 1996a), at the expense of users' cognitive resources (Belkin et al., 2000). 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 are genuinely interest in 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 (Belkin et al., 2000).

On the other hand, implicit feedback techniques tend to collect information on user search behaviour in a more 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 (Morita and Shinoda, 1994; Konstan et al., 1997; Seo and Zhang, 2000) have been all treated as indicators of relevance, despite the lack of sufficient evidence to support their effectiveness (Nichols, 1997). From findings provided by Kelly and Belkin (2001), Kelly and Belkin (2002), and Kelly and Teevan (2003), it is made evident that several reliability issues arise when attempting to infer topical relevance from observable search behaviours, simply because what can be observed does not necessarily correspond to the underlying intention.

As argued in Section §2.4.3, both categories of feedback techniques determine document relevance with respect to the cognitive and situational levels of interaction between the user and the system (Saracevic, 1975). 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. Recent studies suggest that 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 (Damasio, 1994; Reeves and Nass, 1996; Pfister and Böhm, 2008). This suggests that there is a need to reconsider RF with respect to what occurs on the affective level of interaction.

In an earlier study, Kuhlthau (1991) proposed a six-stage model of the ISP based on observations on the search behaviour of high school students. Kuhlthau's findings indicate that the search process is an integration of three dimensions 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 behaviours was delivered by Nahl and Tenopir (1996), Nahl (1998b), and Nahl (2004).

However, limited work has been done with respect to the role of topical relevance as emotional stimuli and its impact on user affective behaviour. Lopatovska and Mokros (2008) 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

Mooney et al. (2006) 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 physiological metrics (galvanic skin response, skin temperature, etc.). The study provides some initial evidence that support the use of biometrics in the latter context.

The study presented in this chapter serves as a starting point for the exploration of the role of emotions in the information seeking process and the impact of task difficulty on user 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 analyse search behaviour at relevance assessment level. The research hypothesis examined here is:

H$_1$**:** Users will experience the same affective states for different levels of task difficulty.

The accumulated evidence suggests that user emotions not only interweave with different physiological, psychological and cognitive processes but also form distinctive patterns, according to specific task, and according to specific user.

4.2 Experimental Methodology

For more details the reader is referred to Section §3.6.1.

4.2.1 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), thus improving or decreasing the chance of finding relevant documents. 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, measurements of the perceived task difficulty, task complexity, information need ambiguity, and other, were also taken.

4.2.2 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.

4.2.3 Apparatus

For this experiment I 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 it was made to appear as inactive by exposing a disconnected power cable that apparently belonged to it.

Logging Software

The desktop computer was equipped with *BB FlashBack*[1] screen recorder that captured participants' desktop actions, without its operation being noticed. Information such as URLs visited, start, finish and elapsed times for interactions, keystrokes and clicks were recorded and stored in a data file located on the desktop computer. To capture participants' facial expressions I used the default 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 recognition system) emotions and saved the scores in 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).

Questionnaires

The participants completed an Entry Questionnaire at the beginning of the study that gathered background and demographic information, as well as previous computer and searching experience. The information obtained from the Entry Questionnaire was used to characterise participants, 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) (Scherer, 2001).

The GAQ was 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

[1] http://www.bbsoftware.co.uk

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 Post-search Questionnaire were forced-choice type, with the exception of a single question that requested a written description. This description asked for the event that elicited the emotional episode, as well as details regarding the consequences it had for the participant. Out of the 35 questions of the GAQ only 18 were used (4-9, 18-23, 25, 29 and 31-34), and retained the structure of categories ii, iii, v, vii and viii in the Post-search Questionnaire. By decomposing the search process to a set of parameters and addressing separately I was able to identify how the different levels of the independent variables affected them. At the end of the study, the participants completed an Exit Questionnaire that gathered information about the perceived task and information need ambiguity, as well as their view of the importance of affective feedback, with respect to usability and ethical issues.

Search Interface

For the completion of the search tasks I employed Indri, an open source search engine from the Lemur project[2]. Indri is a flexible and reliable tool that provides its own complete structural query language and search environment. The interface was modified to appear as one of the popular search engines, 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 was encountered was the complexity of the query language structure.

Test Collection & Search Tasks

For indexing I 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 (Bailey, Craswell and Hawking, 2003). According to Borlund (2000), TREC topics and simulated information need situations share a similar structure, which consists of a number of sections. However,

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

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 subsequent search queries.

In this study, even though the original content of the TREC topics was retained, it was presented using the structural framework of the simulated information need situations. Short *cover stories* were introduced to the participants that allowed the description of the situation, the context, and the information problem to be solved, thus facilitating a better understanding of the search objective (Borlund, 2000). In addition, a layer of realism was added to the search tasks, while preserving well-defined relevance criteria (as the latter are specified by the TREC topic description). Considering the criterion for defining task difficulty, I formulated two different scenarios for each level and allowed the participants to complete the one they considered more interesting.

Facial Expression Analysis Software

In this work I employed eMotion (Valenti et al., 2007), a facial expression recognition system of reasonably robust performance and accuracy across all individuals that applies the discrete-categories approach. The primary goal was to relate participants' facial expressions to the actions taken and the documents assessed. The video recordings were edited using Adobe Premiere Pro CS3. The beginning and ending sections of each recording were trimmed off to isolate the parts of the videos that showed the participants working actively on their search tasks. Those parts were afterwards synchronised with the screen recordings and a *picture-in-picture* effect was applied, followed by manual annotation of each session. For more details the reader is referred to Section §3.6.2.

4.2.4 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 orally and through a Consent Form. After completing an Entry Questionnaire, the session proceeded with a brief training on the use of the search interface. To ensure that the participants' faces would be visible on the web camera they were advised to keep a proper posture, by indicating health and safety measures.

Each participant completed three search tasks in total ("T_1: easy", "T_2: very difficult" and "T_3: practically impossible"). For every task they were suggested two scenarios of similar difficulty and were asked to proceed with the one they preferred the most. Each scenario description provided them with well-defined criteria for document relevance. To negate the order effects the task distribution was counterbalanced 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.

4.3 Results

This section presents the experimental results of this study, based on 72 searches carried out by 24 participants. Questionnaire data were collected 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 my 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 the Bonferroni correction was applied, and so all effects are reported at a .016 level of significance. In addition, performance-related data were gathered 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).

4.3.1 Tasks

Table 4.1 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.

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 4.1: 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

Table 4.2: Descriptive statistics on task aspects

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 was 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 4.1 also shows participant's subjective assessment on the complexity and ambiguity of the three tasks. Friedman's ANOVA test revealed a significant difference for task complexity, but does not indicate the same for task ambiguity. The Wilcoxon tests showed 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, $p < .016$).

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 4.3: Descriptive statistics on search process aspects

Task	**Unpleasantness of stimuli**		**Intensity of emotion**		**Masking emotions**	
	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

Table 4.4: Descriptive statistics on emotional experience

The means and standard deviations for participants' understanding of their information need, as well as the perceived easiness of formulating appropriate query statements, are presented in Table 4.2. 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 showed that the difference among the three tasks is significant for both information need clarity (χ^2(3, N = 24) = 10.314, $p < .05$) and easiness of query formulation (χ^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, $p < .016$).

4.3.2 Search Process

Similarly to the task, I examined the effect of the independent variable to the search process. Table 4.3 shows the means and standard deviations of participants' subjective assessment on search process difficulty, interest and fatigue. The Friedman's ANOVA test showed that search difficulty differed significantly across all tasks (χ^2(3,*N* = 24) = 26.690, p < .05). However, the post-hoc tests showed 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 (χ^2(3, *N* = 24) = 9.896, p < .05). The Wilcoxon tests revealed 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 appeared to differ significantly (χ^2(3, *N* = 23) = 14.986, p < .05). The Wilcoxon tests indicated 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).

4.3.3 Emotional Experience

To evaluate the progression of the emotional patterns across the three tasks the participants were asked to self-assess any emotional episodes they had experienced during the study. Table 4.4 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. A significant difference was found only for the unpleasantness of the stimuli across the different conditions (χ^2(3, *N* = 24) = 14.364, p < .05). The Wilcoxon Signed-Ranked tests that followed up this finding revealed 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 4.1 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. The second pie chart shows a different distribution, with irritation being reported by half of the participants as the most intense emotion for the second task (T_2). Other emotions such as anxiety, anger and happiness were also reported, 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 (χ^2(2, *N* = 24) = 8.33, p < .05) and happiness (χ^2(2, *N* = 24) =11.4, p < .05), across the three tasks.

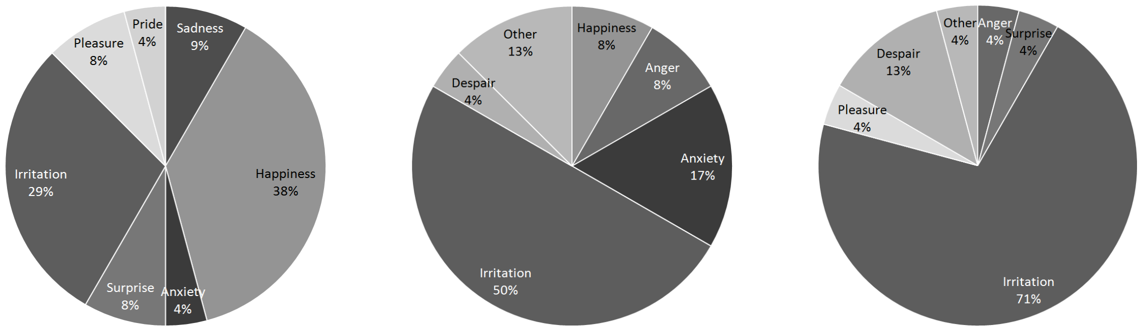

Figure 4.1: Distribution of emotions for Tasks 1-3 (order of appearance: left to right)

4.3.4 Automatic Facial Expression Analysis

In this section I present the preliminary results of the facial expression analysis that was performed using eMotion (Sebe et al., 2007). Each session was processed separately and the data were stored in a log file that was labeled after the participant's unique ID, task number and order number. The log data display, for each key-frame of the video recordings, the probability of the detected facial expression corresponding to any of the seven basic emotional categories that eMotion can recognise (a higher percentage score corresponds to greater confidence in the classification of the detected facial expression, and to higher intensity).

For each log file the number of key-frames, per emotion, that received a probability greater than .90 were counted. The threshold was deliberately set to this high value to exclude emotions that were detected with low probability scores; thus reducing the noise and uncertainty in the data. These scores were then divided with the total number of key-frames of the video sequence, to normalise its contribution to the average values across all videos and per task. The bar chart in Figure 4.2 shows the average values of the aggregates across all participants, for tasks T_1, T_2 and T_3.

In addition, the log data from a random participant were selected and examined to acquire a micro view of the affective patterns that emerge across tasks T_1, T_2 and T_3. The bar chart in Figure 4.3 illustrates the frequency of the seven basic emotions as the latter were determined by eMotion (the scores were again filtered using a .90 threshold). The dissimilar distribution of scores makes evident the emotional blend that characterises 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.

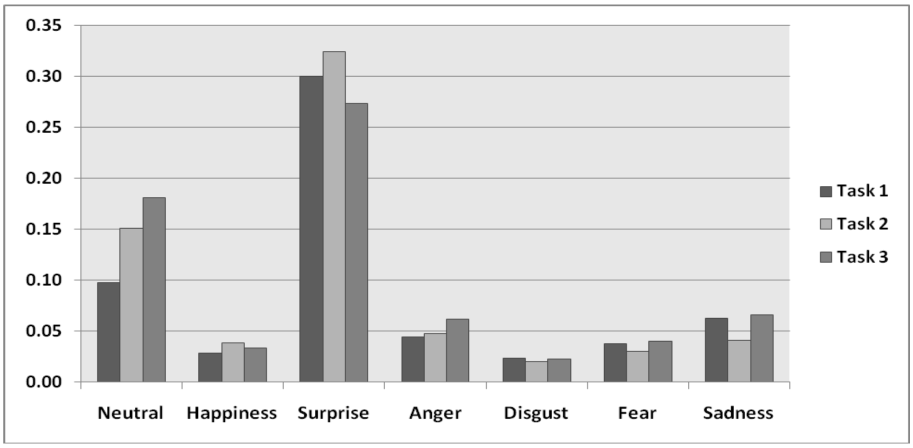

Figure 4.2: Average scores of detected emotions, across all participants, for Tasks 1-3

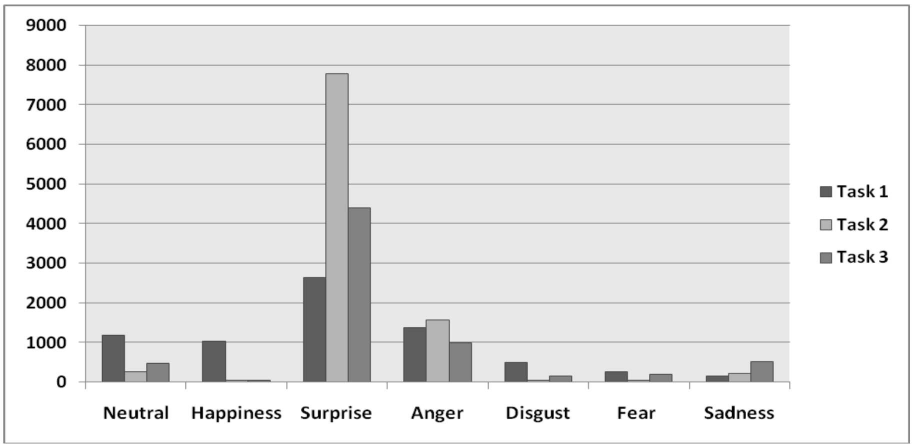

Figure 4.3: Aggregated scores of detected emotions from a random sample, for Tasks 1-3

4.4 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 always perceived as such.

We 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, the 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 respect to 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 4.3 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 compared to 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 4.1 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. I speculate 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 behaviour of the seven basic emotional categories, in terms of frequency, as they are illustrated in Figure 4.2. 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 results 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). Perhaps the low frequency scores of some emotions are suggesting that the latter might make better feedback indicators, compared to other categories that exhibit higher scores. I 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 the follow-up analysis.

Finally, the bar chart in Figure 4.3 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 a somewhat larger subset, it is not possible to generalise whatever conclusions are drawn 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.

4.5 Summary

In this chapter I presented an exploratory user study that involved 24 participants and allowed for the collection of multi-modal data. Several interesting conclusions can be drawn here. 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 significant events and breakdowns that are correlated with changes in the users' knowledge state and information need. Moreover, the findings reveal that participants' emotions progressively transit from positive to negative valence, as the degree of task difficulty increases, which negates my research hypothesis. It also suggests that 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. Despite its limitations, the presented study is the first step into a new and unexplored domain and a contribution to the exploration of the role of emotions in the ISP. Finally, it introduces a new approach to the detection and quantification of affective information, in an attempt to reconsider RF on a cognitive, as well as affective level.

Chapter 5

User Study 2: Enriching a Multimodal Recommender System with Affective Features

Preamble

Recommender systems have been systematically applied in industry and academia to help users cope with information uncertainty. However, given the multiplicity of the preferences and needs it has been shown that no approach is suitable for all users, in all situations. Thus, it is believed that an effective recommender system should incorporate a variety of techniques and features, to offer valuable recommendations and enhance the search experience. In this chapter I propose a novel video search interface that employs a multi-modal recommender system, which can predict the topical relevance of viewed results. The proposed system accounts for interaction data, contextual information, as well as users' affective responses, and exploits these information channels to provide meaningful recommendations of unseen videos. The experimental findings reveal that affective feedback can complement existing feedback techniques and it is a promising way to improve the quality of recommendations.

5.1 Introduction

During the past two decades the explosive growth rate of the world wide web has led to a profound increase in the availability of online multimedia content (Odlyzko, 2003). Considering the scale of this growth and the information overload problem that it introduced, it becomes evident that the average user often faces a challenge when attempting to locate online items of interest or relevance. To address these issues different techniques have been developed that aim to improve some of the facets of IR, such as indexing, searching and information filtering. However, while they often provide a varying level of support, they are not always tailored to the specific user and the specific situation.

In recent years recommender systems have emerged as a potential solution to the problem of information overload. Recommender systems are a personalised information filtering technology, designed to assist users in locating items of interest, by providing useful recommendations (Han and Karypis, 2005). They have been applied successfully in a number of different applications to improve the quality of web services. Such examples include Amazon.com, for recommending books, CDs and other products (Linden, Smith and York, 2003), MovieLens, for recommending movies (Miller, Albert, Lam, Konstan and Riedl, 2003), and VERSIFI Technologies, for recommending news articles (Billsus, Brunk, Evans, Gladish and Pazzani, 2002). Recommender systems adopt various profiling techniques to collect information about the interaction history, which they eventually integrate into user profiles. The data retained inside the profiles are regarded as indicative of user interests (Adomavicius and Member-Tuzhilin, 2005), and, usually, refer to information such as age, gender, place of birth, preferences, needs, etc. Based on the internal form of representation of the user information, the latter profiles can be categorised into single-faceted and multi-faceted.

User profiling consists of three stages: (i) relevance feedback, (ii) feature selection, and (iii) updating of profile. The feedback cycle is a necessary practice, since users are sometimes guided by a vague information need, which they cannot easily express in terms of keywords, or relate to unseen information items (Harman, 1992). The value of relevance assessments lies in the progressive disambiguation of that need and it is usually achieved through the application of different feedback techniques. These techniques range from explicit to implicit and help determine the relevance of the retrieved items. However, they often do so by determining relevance with respect to the cognitive and situational levels of interaction, failing to acknowledge the importance of intentions, motivations and feelings in cognition and decision-making (Damasio, 1994; Pfister and Böhm, 2008).

In this chapter I propose a novel video search interface that applies real-time facial expression analysis to aggregate information on user affective behaviour. Furthermore, I present a way of analysing that information to determine the topical relevance of perused videos and, eventually, enrich the user profiles. The value of the interface lies in the combination of different modules (facial expression recognition, recommender system, etc.), the integration of sensory data and, moreover, the application of information fusion. The work presented here contributes also to the exploration of the role of emotions in the search process, by highlighting some of the factors that can influence user affective behaviour, and provides answers to the following research hypotheses:

H$_1$**:** Users' affective responses are consistent across different types of stimuli (search process, the viewed content).

H$_2$**:** The integration of affective features, deriving from automatic facial expression analysis, in user profiling can improve the performance of a recommender system.

H$_3$**:** Affective feedback can effectively complement existing feedback techniques, such as click-throughs.

5.2 Recommendation Model

The interests of each individual is stored in a profile, which is generated during registration time. User interests are dynamic in nature and can change over time. It is, therefore, important to have a system that can accommodate such changes. Moreover, the efficiency of the system is dependent on the accuracy of captured interests; thus, it is important to break-down interaction into several phases, allowing us to develop a better understanding of these interests and their changes. The first phase of capturing user interests is during query submission. At this point the system is able to perform recommendations, using the terms that appear in the search query. The profile is updated each time the user formulates a new query.

The second phase occurs during click-through action. This action is regarded as an implicit indicator of interest towards the selected items. The items' meta-data is used as a source of information for updating the user profile. These two steps consist the baseline user-profiling technique. In the enriched user-profiling technique the captured facial expressions are treated as an additional source of implicit feedback, which is used to update the user profile. The perused items' meta-data are also used as an additional source of information, along with the positive and negative feedback obtained from the facial expression analysis. A feedback is regarded as positive if the user finds at least one frame interesting during the time he/she is watching a video.

Each of these actions has its own degree of significance. The search query is considered to be the least significant, since users often have problems expressing their initial information need. On the contrary, click-throughs are considered a more reliable source of feedback, since users have the opportunity to go through the meta-data and decide whether to view the video or not. Finally, any feedback deriving from the facial expression analysis is regarded as the most significant, because it is generated while the users are watching the actual content of the video. After each feedback cycle the user profile is updated, following the multi-modal approach proposed by Cetintemel, Franklin and Giles (2000).

5.3 Experimental Methodology

For more details the reader is referred to Section §3.6.1.

5.3.1 Design

This study used a repeated-measures design. There were three independent variables, namely: task domain (with two levels: "learning" and "entertainment"), task scope (with two levels: "broad" and "focused"), and recommendation system (with two levels: "RS1: baseline" and "RS2: multi-modal"). The task domain levels were controlled by assigning topics with the appropriate context, while the task scope levels were controlled by introducing either well-defined or less explicit relevance criteria. The recommendation system levels were manipulated by employing a different user profiling technique. The baseline version of the proposed system applied a profiling technique which integrated information that derived only from participants' actions (meta-data & click-throughs). The multi-modal version integrated affective information (participants' facial expressions) on top of the interaction data that was captured. The dependent variables were the system's performance, as it was perceived by the participants, as well as their emotional experience with respect to the search process and the viewed content.

5.3.2 Participants

Twenty-four participants of mixed ethnicity and educational background (3 Ph.D. students, 12 MSc students, 4 BSc students and 5 other) applied for the study through a campus-wide ad. They were all familiar with the English language (4 native, 12 advanced, 4 intermediate and 4 beginner speakers). Of the 24, 13 were male and 11 were female. All participants were between the ages of 19 and 37 and free from any obvious physical or sensory impairment.

5.3.3 Apparatus

For this experiment I used two desktop computers equipped with conventional keyboard and mouse. The first computer acted as the server, which hosted the recommender system, the Support Vector Machine (SVM) model, the facial expression recognition system (eMotion) and the video recording software. The second computer acted as the client and was used to provide access to the search interface. In addition, participants' desktop actions were logged using a custom-made script, which recorded information such as starting, finishing and elapsed times for interactions, and click-throughs. A "Live! Cam Optia AF" web camera, with a 2.0 megapixels sensor, and a "Logicool Qcam", with a 1.3 megapixels sensor, were also mounted on top of the client's screen. The cameras were used for recording the participants' expressions and apply real-time facial expression analysis.

Questionnaires

The participants completed an Entry Questionnaire at the beginning of the study, which gathered background and demographic information, as well as previous experience with multimedia and online searching. The information obtained from it was used to characterise participants, but not in subsequent analysis. A Post-Search Questionnaire was also administered after each task, to elicit participants' viewpoint on certain aspects of the search process. The questionnaire was an adaptation of the Geneva Appraisal Questionnaire (GAQ) (Scherer, 2001). Its purpose is to assess, through recall and verbal report, the results of participants' 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 elicited the emotional episode, in addition to details regarding what has happened and the consequences it had for the participant. Out of the 35 questions of GAQ we used only 8 (4-6, 25, 29 and 32-34) that were relevant to the context of this study (for more details refer to Section §4.2.3). Finally, an Exit Questionnaire was introduced at the end of the study.

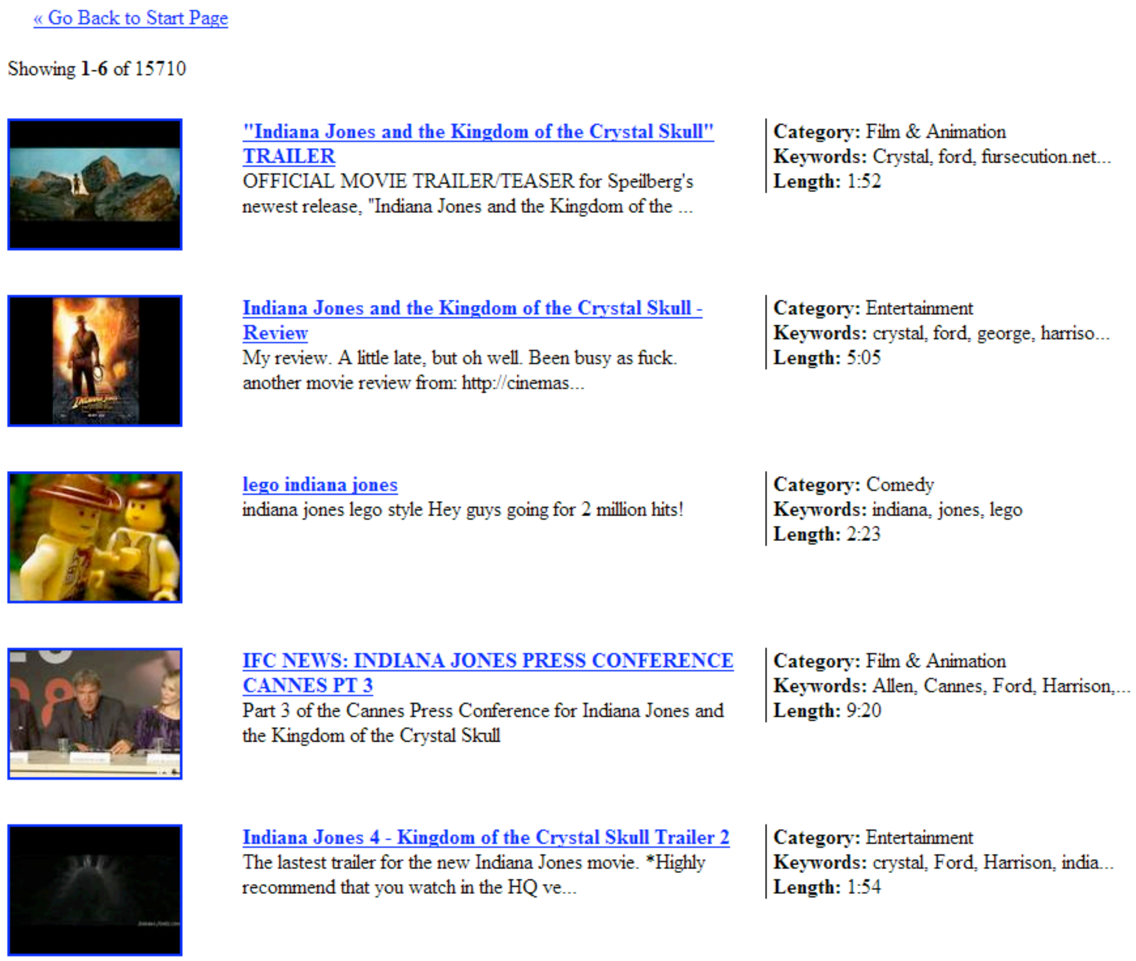

Figure 5.1: The results page (second layer)

Search Tasks

A set of search tasks that differed in their domain and scope was formulated. All topics were manually performed prior to the experiment, to ensure the availability of relevant videos in the YouTube collection. The criteria for selecting the search tasks were that there should be enough relevant documents for each topic, to allow for the accumulation of sufficient data for the user profiles. The tasks were presented using the structural framework of the simulated information need situations (Borlund, 2000). By doing so, a better understanding of the task was facilitated and the participants' motivation was increased. For every search task the participants had the possibility of selecting among a predefined list of options the sub-topic of their choice.

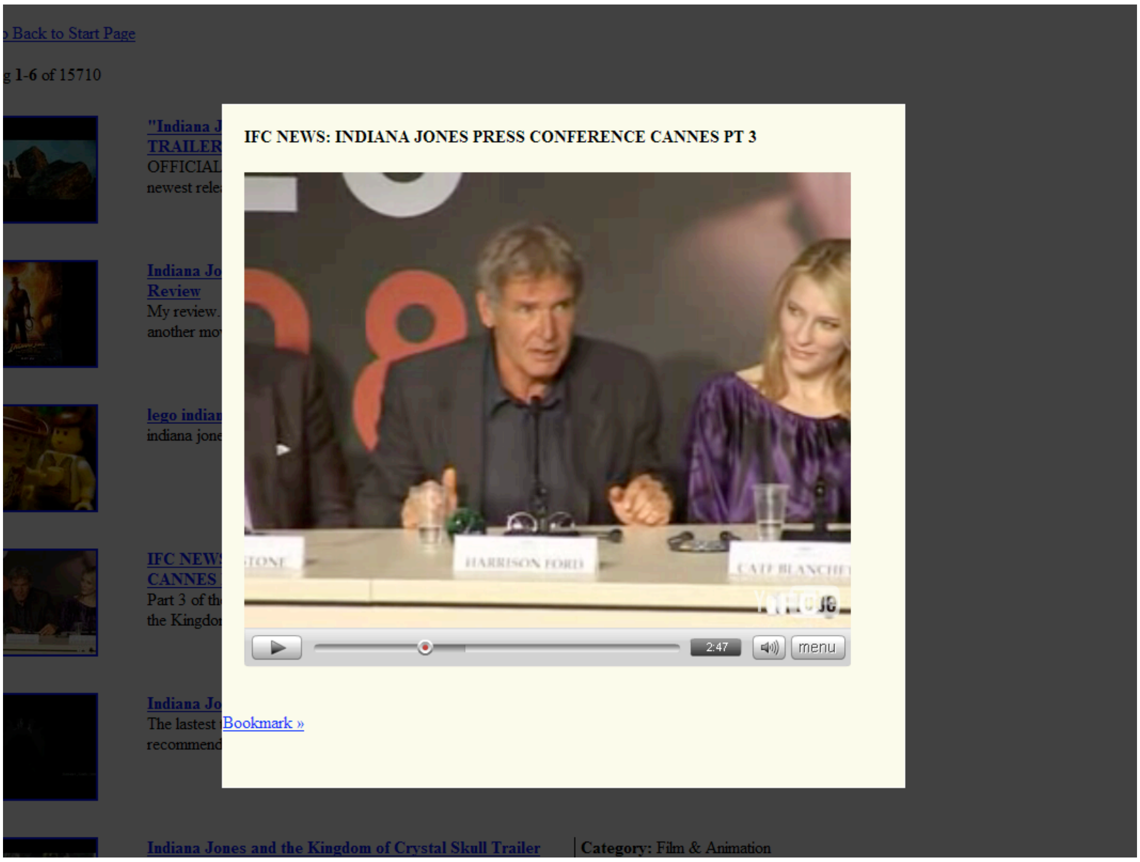

Figure 5.2: Browsing a video (third layer)

Search Interface

For the completion of the search tasks a customised video search interface was employed. The interface worked on top of the YouTube search engine and was designed to resemble its basic layout, while retaining a minimum number of graphical elements. Each result was represented by a thumbnail, a short description and some meta-information (category, associated keywords and duration). The interface used the YouTube API to retrieve video clips from the YouTube collection and present them in their original order (Figure 5.1). The participants had the option to either perform a focused search, by formulating and submitting a query, or by looking into the available video categories for relevant clips.

The architecture of the video search interface consists of three different layers. The first layer is dedicated to support any interaction that occur at the early stages of searching, such as query formulation and search execution. At this layer the participants could submit their queries or explore the predefined video categories. Any output generated by that interaction (whether originating from a user query or the selection of a video cat-

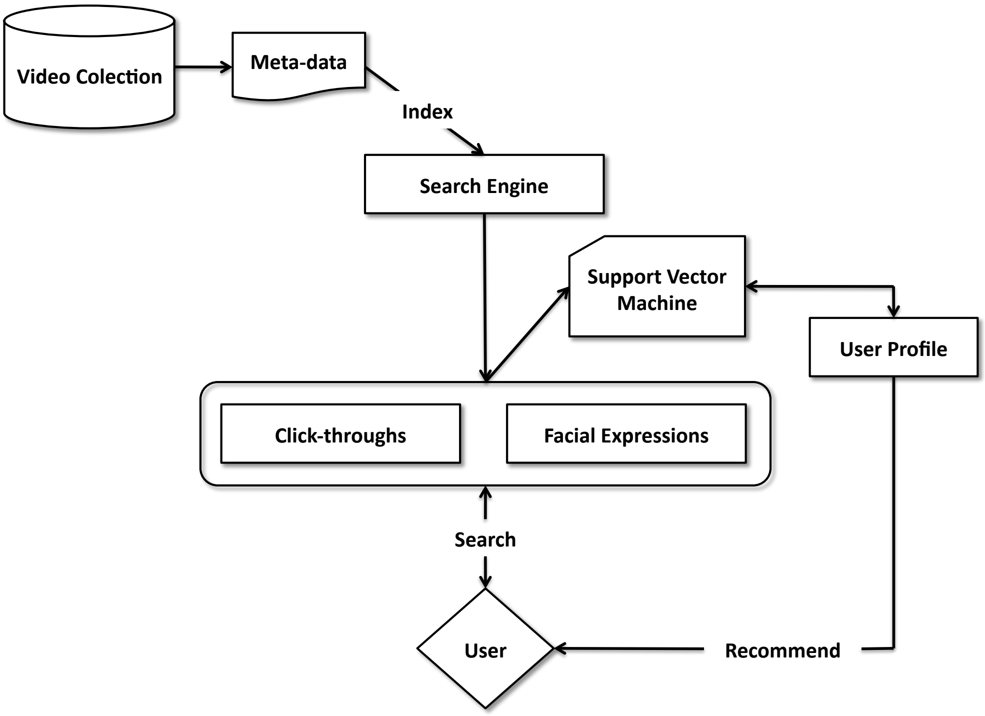

Figure 5.3: System architecture

egory) is presented in the second layer. From there, the participants could easily select and preview any of the retrieved clips. The content of a clip is shown on a separate panel, in the foreground, which corresponds to the third layer of the proposed system (Figure 5.2). The main reason behind this layered architecture was to isolate the viewed content from all possible distractions that reside on the desktop screen; therefore, establishing reasonable ground truth that allowed us to relate the recorded facial expressions to the source of stimuli (in this case, the perused video clip). Upon viewing the clip, the participants had to explicitly indicate: (i) the degree of relevancy of the video, and (ii) the emotional impact of the video content.

SVM Model

A two-layer, hierarchical SVM model was trained to discriminate between two categories of videos (relevant, irrelevant) by analysing facial expression data. The model was trained using a radial basis function (RBF) kernel, which, among the basic four SVM kernels (linear, polynomial, radial basis function, sigmoid), was considered as a reasonable first choice. The RBF kernel can nonlinearly map data into a higher dimensional space, unlike the linear kernel that can be applied successfully only when the relation between the class labels and the attributes is nonlinear (Hsu, Chang and Lin, 2003). In addition, the sigmoid kernel behaves like an RBF, for a certain set of parameters (Lin and Lin, 2003). Moreover, the RBF kernel is preferable, since it encounters less numerical difficulties and has a limited number of hyper-parameters.

The ground truth was obtained by classifying relevant vs. irrelevant expressions in the annotated data set that was acquired from Arapakis et al. (2008). It is acknowledged that the training data derived from a document retrieval experiment and was, therefore, not portraying very accurately the conditions that were encountered during the video retrieval tasks. However, it was the only available annotated data set that was available at that point.

The SVM model consists of 10 weak classifiers, each trained on a different instance of the training set. The whole training set was predicted once, and the output of each weak classifier (support vectors) was used to train the meta-classifier. To deal with the imbalanced set, the SVM model was trained by randomly sampling half of the relevant key-frames and one-fifth of the irrelevant key-frames. Furthermore, each time a key-frame appeared in both categories it was labeled as relevant. This hierarchical framework improved the original accuracy from 78% to 89%.

Facial Expression Analysis Software

In this work I employed eMotion (Valenti et al., 2007), a facial expression recognition system of reasonably robust performance and accuracy across all individuals that applies the discrete-categories approach. The primary goal was to relate participants' facial expressions to the topical relevance of perused videos and fuse that information with interaction data. Every time a clip was perused eMotion applied facial expression analysis for every key-frame captured by the camera, during that interval. It then communicated to a pre-defined port the results of the emotion classification, along with the corresponding motion units, as a stream of sensory data. The streamed data was then forwarded to the SVM model and, depending on the outcome of the classification, the perused video was labeled as either relevant or irrelevant. In the first case, the recommender system attempted to retrieve more similar results, using the meta-information of the perused video clip. For additional details regarding eMotion the reader is referred to Section §3.6.2.

5.3.4 Procedure

The user study was carried out the following way: The formal meeting with the participants took place in the office of the researcher. At the beginning of each session the participants were informed about the conditions of the experiment, both orally and through a Consent Form, and then completed an Entry Questionnaire. A brief training followed, which explained the basic functions of the search interface environment and the terms of interaction. Also, to ensure that participants' faces would be visible to the camera at all times we encouraged them to keep a proper posture, by indicating health and safety measures.

Every participant completed two search tasks in total. For each search task they were given a short cover story which introduced them to an artificial situation, thus facilitating the formulation of better-defined relevance criteria. To negate the order effects the task distribution was counterbalanced by using a Latin Squares design. The participants were asked every time to bookmark as many relevant videos as possible and were given 15 minutes to complete the task, during which they were left unattended to work. At the end of each task the participants had to complete the first part of a Post-search Questionnaire and afterwards evaluate a set of recommended videos, which were selected using one of the recommendation strategies. After that they were asked to complete the second part of the Post-search Questionnaire.

An Exit Questionnaire was also administered at the end of each session, along with the receipt of payment. Finally, the participants were asked to sign a Payment Form, prior to receiving the participation fee of £12.

5.4 Results

This section presents the experimental results of this study, based on 48 sessions that were carried out by 24 participants. Questionnaire data was collected on three aspects of the information seeking process, namely: (i) perceived relevance of recommendations, (ii) emotional experience with respect to the search process, and (iii) emotional experience with respect to the viewed content. A 5-point Likert scale was used in all questionnaires. Questions that ask for participant rating on a unipolar dimension have the positive concept corresponding to the value of 1 (on a scale of 1-5) and the negative concept corresponding to the value of 5. Questions that ask for participant rating on a scale of 1-5 represent in my analysis stronger perception with high scores and weaker perception with low scores.

Friedman's ANOVA and Wilcoxon Signed-Ranked test were used to establish the statistical significance ($p < .05$) of the differences observed among the four types of tasks ("Learning - Broad", "Learning - Focused", "Entertainment - Broad" and "Entertainment - Focused"). Pearson's Chi-Square test and the Dependent t-test were applied in the analysis of emotion variance (between search process and viewed content) and the performance of the recommender systems. To take an appropriate control of Type I errors I applied a Bonferroni correction, and so all effects are reported at a .0125 level of significance.

5.4.1 Recommender Systems

This section presents the results from the evaluation of the recommender systems performance, as it was determined by the participants' ratings. The main and interaction effect of the independent variables are examined, again with respect to reported relevance. To distinguish the first half (search task) from the second half (evaluation or recommended videos) of each session, I refer to the former as "Initial" and to the latter as "Recommendation". The recommender systems are also categorised as "Baseline" and "Multi-modal".

Table 5.1: Average rating of recommended videos

	Baseline	Multi-modal	Total
Overall	1.7 (1.4)	**2.0** (1.5)	1.8 (1.4)
Domain: Learning	1.8 (1.4)	**2.3** (1.6)	2.0 (1.5)
Domain: Entertain.	1.6 (1.3)	1.7 (1.3)	1.6 (1.3)
Scope: Broad	1.8 (1.5)	**2.2** (1.6)	2.0 (1.6)
Scope: Focus	1.6 (1.2)	1.8 (1.4)	1.7 (1.3)

Bold: Statistically significant at $p \leq .05$.

Main Effect

Table 5.1 shows the means and standard deviations (in brackets) of participants' ratings for the two recommendation systems. The second row shows the overall performance of the two systems. As can be seen, participants gave a higher rating to the videos recommended by the multi-modal system when compared to the baseline system. The Mann-Whitney Test showed that the difference is significant (W = 28791.5, p = 0.020). Note that the independent test was used, since participants made several ratings within individual blocks, although the experiment was a within-subject design.

Domain & Scope Effect

The effect of tasks on the performance was also of interest: task domains and task scope. First, I split the rating data based on the blocks of domains or scopes. Then the Mann-

Whitney Test was applied to individual blocks. The results are presented in rows 3 to 6 of Table 5.1. As can be seen, the difference between the two systems was found significant for the "Learning" set of the task domain (W = 7297, p = 0.006), and "Broad" set of the task scope (W = 7550.5, p = 0.015).

Two-way ANOVA tests were also applied by using systems and task domains as independent variables. The results show that both main effects are significant but no interaction effect was found. The same test was repeated for the system type and task scopes. The results were similar: significant main effects without interaction effect. Therefore, more research is needed to determine the effect of tasks on the system performance, although some supporting evidence has been found.

Domain	**Unpleasantness of stimuli**		**Intensity of emotion**		**Masking of emotions**	
	M	SD	M	SD	M	SD
Learning	2	1.0488	3.1905	1.1233	2.0476	1.0235
Entertainment	1.7619	.9952	3.3333	.9661	1.8571	1.0142

Table 5.2: Descriptive statistics on emotional experience (domain effect)

Scope	**Unpleasantness of stimuli**		**Intensity of emotion**		**Masking of emotions**	
	M	SD	M	SD	M	SD
Broad	2	1	3.1429	1.1526	2.0952	.9952
Focused	1.619	.9207	3.1429	3.1429	1.8571	1.0623

Table 5.3: Descriptive statistics on emotional experience (scope effect)

Task	**Unpleasantness of stimuli**		**Intensity of emotion**		**Masking of emotions**	
	M	SD	M	SD	M	SD
Learn. - Broad	2.2222	1.0929	3.3333	1.3229	2	.866
Learn. - Focused	1.7778	1.0929	3.4444	.7265	2.2222	1.3017
Entert. - Broad	1.7778	.9718	3.3333	.866	2.1111	1.1667
Entert. - Focused	1.3333	.7071	3.3333	.7071	1.5556	.8819

Table 5.4: Descriptive statistics on emotional experience (interaction effect)

5.4.2 Emotional Experience

To evaluate the progression of the emotion patterns across the tasks, the participants were asked to self-assess the emotional episodes they experienced during the study. Tables 5.2, 5.3 and 5.4 show 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 emotions, as well as the amount of effort that the participants put to control or mask their emotional expressions. Friedman's ANOVA was applied to evaluate the effect of the independent variables to participants' emotion behaviour but it did not reveal any statistically significant differences. I also examined the main effect of the task domain and task scope on the same variables. Again, the Friedman's ANOVA test did not reveal any significant differences between any of the individual tasks.

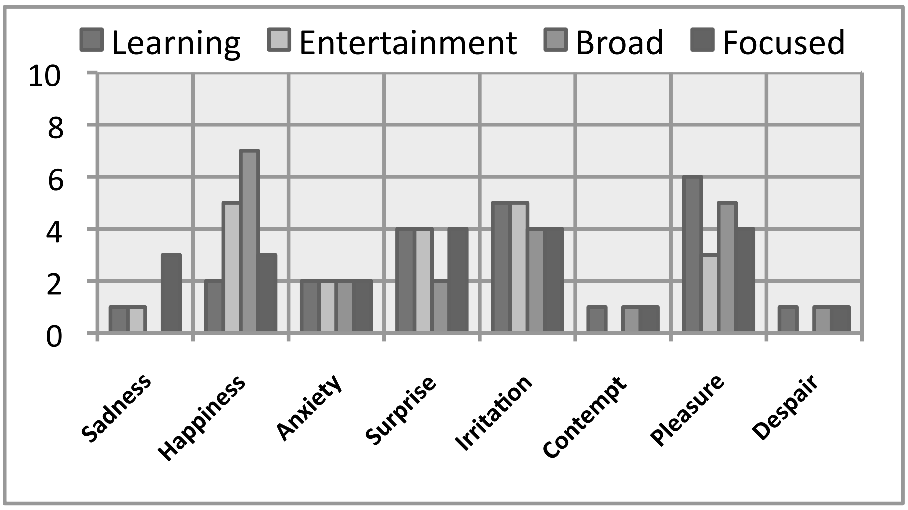

Figure 5.4: Reported emotions (domain & scope effect)

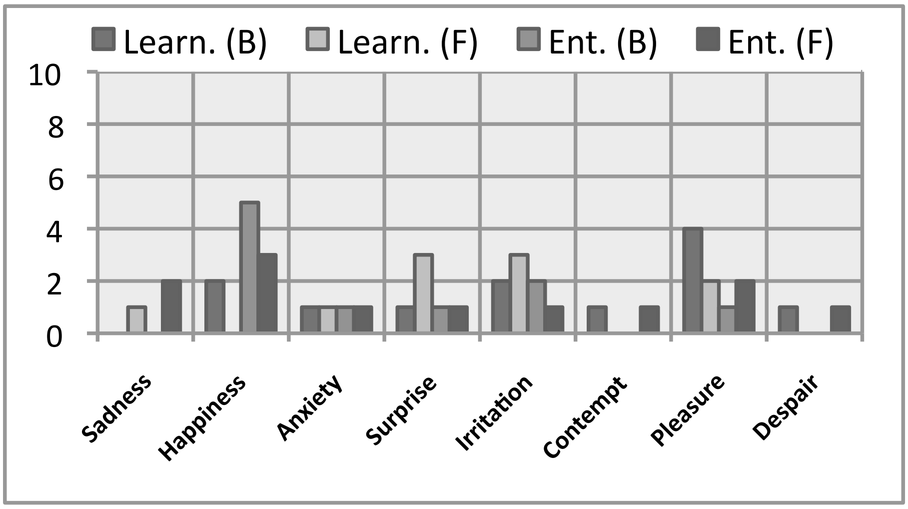

Figure 5.5: Reported emotions (interaction effect)

Search Process

A major goal of this study was to examine the occurrence of emotions during multimedia retrieval. The column chart in Figure 5.4 illustrates the effect of the task domain on the emotions of the participants, within the context of the search process. It appears that pleasure, irritation and surprise were the most intense emotions for tasks in the learning domain, followed by happiness, anxiety, contempt, despair and sadness. The Entertainment domain follows a similar pattern, with happiness, irritation and surprise being reported as the dominant emotions, followed by pleasure, anxiety and sadness. The column chart also depicts the effect of the task, again with respect to the search process. It is evident that happiness and pleasure were the dominant emotions for the tasks with broad scope, followed by irritation, anxiety, surprise, despair and contempt. A more balanced distribution characterises the tasks with focused scope, which appear to have pleasure, happiness, irritation and surprise as equally dominant, followed by sadness, anxiety, despair and contempt. Pearson's Chi-Square test was applied to both domain and scope categories, but it did not reveal an associated relationship between them.

The column chart in Figure 5.5 illustrates the distribution of the most intense emotions with respect to the search. We can see that for the first task category (Learning - Broad), pleasure, happiness and irritation were the most intense, among all other reported emotions, followed by despair, anxiety, contempt and surprise. The second task category (Learning - Focused) shows a different distribution, with irritation and surprise being reported by almost one-third of the participants as the most intense emotions. Other emotions, such as pleasure, anxiety and sadness were also reported, but at a lesser rate. The third task category (Entertainment - Broad) is dominated mainly by emotions of happiness, followed by irritation, anxiety, surprise and pleasure. Finally, the fourth task category (Entertainment - Focused) has a more balanced emotional blend, with happiness being the only dominant emotion, followed by pleasure, sadness, anxiety, despair, contempt, irritation and surprise. Pearson's Chi-Square test was again applied without revealing an associated relationship between any of the emotions and the task categories.

Viewed Content

The column chart in Figure 5.6 illustrates the effect of the task domain on the emotional experiences of the participants, with respect to the viewed content. For the Learning domain there were many reports of surprise and pleasure, followed by happiness, irritation, fear and sadness at a lesser rate. The Entertainment domain shows a similar pattern, with happiness and pleasure as the dominant emotions, followed by surprise, sadness, disgust and irritation. The tasks with broad scope appear to have happiness and pleasure as the dominant emotions, followed by surprise, fear, irritation, sadness and disgust. A very similar distribution also appears in the tasks with focused scope, with pleasure, happiness being again the dominant emotions, followed by surprise, sadness, irritation and disgust. Pearson's Chi-Square test was applied but it did not reveal an associated relationship between any of the emotions for the task domain or scope.

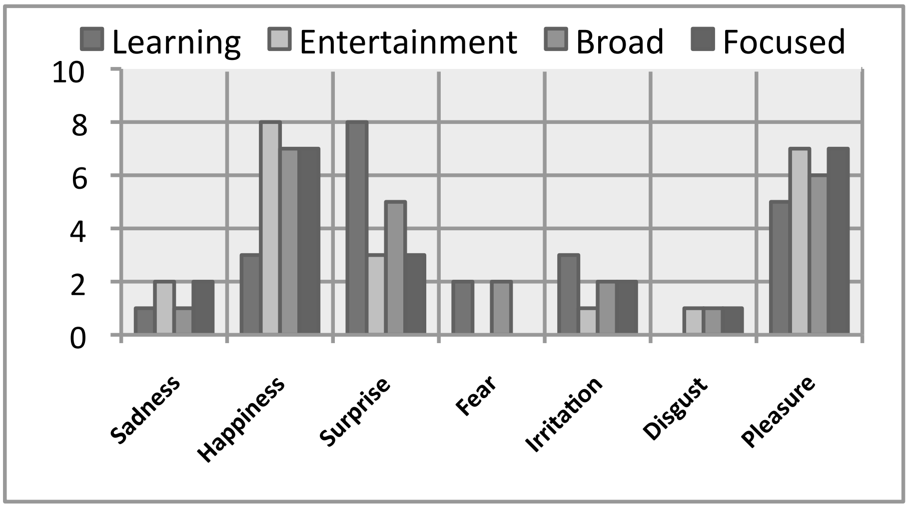

Figure 5.6: Reported emotions (domain & scope effect)

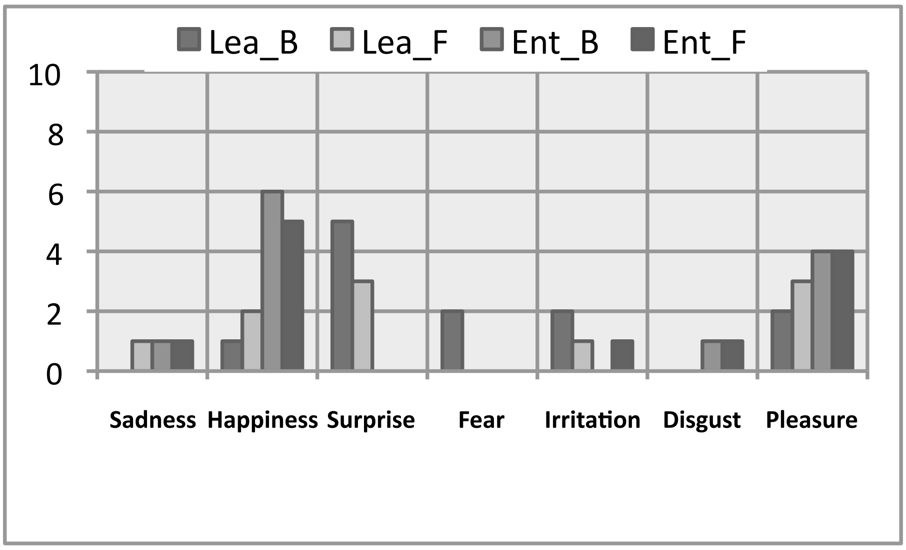

Figure 5.7: Reported emotions (interaction effect)

The column chart in Figure 5.7 shows the distribution of the most intense emotions, with respect to the viewed content. It is evident that for the first task category surprise was the most dominant, among all other reported emotions, followed by pleasure, irritation, fear and happiness. The second task category has a similar blend, with pleasure and surprise being reported by almost one-third of the participants as the most intense emotions. Other emotions, such as happiness, irritation and sadness were also reported at a lesser rate. The third task category is primarily dominated by happiness, which was reported by half of the participants, and followed by pleasure, sadness and disgust. Likewise, the fourth task category appears to have significantly more reports of happiness, followed by pleasure, sadness, irritation, and disgust. Pearson's Chi-Square test was applied but it only revealed an associated relationship between the task category and the emotions of happiness ($\chi^2(1, N = 48) = 4.181, p < .05$) and surprise ($\chi^2(1, N = 48) = 5.839, p < .05$).

Search Task & Viewed Content

Pearson's Chi-Square test was applied but it did not reveal an associated relationship between any of the emotions and the emotional stimuli (search process, viewed content), for all domains (learning, entertainment), scope (broad, focused), as well as their interaction effect.

5.5 Discussion

The post-hoc analysis of the results indicates that both recommender systems facilitated a more effective search by suggesting unseen relevant videos to the participants, with the multi-modal recommender achieving a higher performance. This finding suggests that the performance of profiling was enhanced by the facial expression data that proved to be a better relevance predictor than click-throughs, which validates research hypotheses H_2 and H_3. Both main and interaction effect analysis showed that the participants gave higher ratings for the task domain "Learning" and task scope "Broad". This finding indicates that the multi-modal system was more effective than the baseline when the tasks involved some form of learning or when the tasks involved a wide range of videos.

Overall, the multi-modal recommender system was found to perform better than the baseline (H_2), which accounted only for click-through data, since the differences in the performance of the two systems were found to be statistically significant. The analysis of the results also outlines the benefits of enriched profiling and the use of facial expression data over other implicit feedback indicators (H_3), and supports the design of multi-modal recommender systems. However, it is acknowledged that additional evaluation of the models is necessary, using more sophisticated training techniques and a data set that addresses the conditions of video, rather than document retrieval.

The analysis of the emotional experience of the participants, as this was inferred from the questionnaire data, revealed that very little effort was put to mask their emotional expressions. Therefore, we can reasonably assume that the captured facial expressions were spontaneous and authentic. This behaviour appears to be consistent across the two tasks, regardless of domain or scope of the search task. It also suggests that, from the viewpoint of the participants, the presence of the recording equipment did not affect significantly their emotional behaviour. Moreover, the intensity of the experienced emotions did not vary significantly across the tasks, indicating that the participants experienced emotions of average intensity. This was an anticipated finding, since the employed video collection was not expected to induce strong emotions to the participants. This was a desired condition, since any extreme/strong affective states detected were more likely produced during the appraisal process of the videos' relevance to the participant's information need. Similarly, the unpleasantness of the emotional stimuli was also found consistent across the different tasks, revealing a trend towards positive stimuli.

In relation to the search process, the column graphs in Figures 5.4 - 5.7 did not make evident any emerging emotion patterns, unlike the progression from positive to negative emotions reported by Arapakis et al. (2008). It appears that neither the domain, scope, or the interaction between them, had an effect. The latter finding is supported by the results of the Pearson's Chi-Squared test, which did not indicate any associated relationship between the reported emotions and the independent parameters of the study. Similarly to the emotion patterns of the search process, those that emerged in the context of the viewed videos do not appear to vary significantly between the domain or scope. Again, this finding is supported by the Pearson's Chi-Squared test results, which did not reveal any associated relationship. However, the interaction of the independent variables had an effect on two of the reported emotions. More specifically, the Pearson's Chi-Squared test showed that happiness was significantly more frequently reported in the "Entertainment - Broad" task, compared to the "Learning - Broad" task ($\chi^2(1, N = 48) = 4.181$, $p < .05$), while surprise was significantly less frequently reported in the "Entertainment - Broad/Focused" tasks, compared to the "Learning - Broad" task ($\chi^2(1, N = 48) = 5.839$, $p < .05$).

Looking at the first research hypothesis, the post-hoc analysis did not reveal an associated relationship between the reported emotions and the emotional stimuli (search process, viewed content), for any of the domain or scope categories. Therefore, we do not have enough evidence to refute the null hypothesis ($\boldsymbol{H}_0$: participants' affective responses are not consistent across different types of stimuli). This, however, does not necessarily mean that the null hypothesis is true. It only suggests that there is not sufficient evidence against $\boldsymbol{H}_0$, in favour of $\boldsymbol{H}_1$.

5.6 Summary

In this chapter I studied the behaviour of emotions in the feedback process and tested the application of affective feedback in an operational system. Overall, the study's findings validate both hypotheses H_2 and H_3, namely that affective feedback, as determined from automatic facial expression analysis, can improve the performance of a recommender system and, in addition, complement existing feedback techniques, such as click-throughs. The novel video retrieval system I introduced accounts for user feedback deriving from real-time facial expression analysis. The value of the proposed system lies in the combination of different modules and modalities, as well as the seamless integration of affective elements into user profiling. This approach can facilitate and sustain a different form of feedback, one which accounts for the affective dimension of human-computer interaction. In addition, it offers a novel way to process that information to determine topical the relevance of videos and generate meaningful recommendations. In the next chapter I investigate ways of improving the affective models' performance and evaluate a range of sensory channels, as well as classification techniques.

Chapter 6

User Study 3: Using Affective Feedback as an Implicit Indicator of Topical Relevance

---

Preamble

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

6.1 Introduction

Multimedia search systems face a number of challenges emanating mainly 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 (Smeulders, Worring, Santini, Gupta and Jain, 2000). Effective search techniques are needed to deal with a variety of multimedia data. Though progress has been made, the effectiveness of existing systems is still limited. The gap is further widened when the user is driven by an ill-defined information need, often the result of an anomaly in his/her current state of knowledge (Belkin, 1980). The formulated search queries, which are used by the search system to locate potentially relevant items, produce results that do not address the user needs.

Implicit feedback is considered a useful technique in addressing many of the semantic-related issues. By analysing implicit feedback IR systems can tailor the search criteria to address more effectively users' information needs. However, as argued in Section §2.4.3, existing feedback techniques determine content relevance only with respect to the cognitive and situational levels of interaction, failing to acknowledge the importance of intentions, motivations and feelings in cognition and decision-making. There is evidence that people naturally express emotion to machines and introduce a wide range of social norms and learned behaviours that guide their interactions with, and attitudes toward, interactive systems and information items (Reeves and Nass, 1996; Vinciarelli et al., 2009).

With respect to online search behaviour a number of studies from the field of Library and Information Science (LIS) have provided evidence which suggests that affect can influence several aspects of the search process, such as search strategies (Nahl and Tenopir, 1996), performance (Wang et al., 2000; Nahl, 1998b), and satisfaction (Nahl, 2004). Positive and negative emotions have been associated with satisfactory search results (Tenopir et al., 2008), successful search completion (Bilal and Kirby, 2002) and interest in the process and documents (Kracker, 2002; Lopatovska and Mokros, 2008).

According to McKechnie et al. (2007), affective variables can play an important role in reading-related information behaviour, especially in the domain of everyday life. Information processing, which occurs during the appraisal process of a goal, an event, or an item, can result in a series of changes in the user's cognitive and affective states (Scherer, 2001). I argue that such changes are often expressed through a psycho-physiological mobilisation that is reflected by a series of more or less observable cues, such as facial expressions, body movements, localised changes in the electrodermal activity, variations in the skin temperature, and many more. Since the significance of an event or information item can vary from low to high, depending on the number of goals or needs that are affected by it, so do these peripheral physiological symptoms vary in intensity and duration.

The modelling and integration of such affective features in the feedback cycle could allow search systems to facilitate a more natural and meaningful interaction. In addition, it could improve the quality of the query suggestions, and, potentially, influence other facets of the information seeking process, such as indexing, ranking and recommendation (see §2.4.3). Eventually, it may be that relevance inferences obtained from this kind of models will also provide a more robust and individualised form of feedback, which will allow us to deal effectively with the semantic gap.

In this chapter I explore the role of affective feedback in designing multimedia search systems. I investigate whether topical relevance can be deduced by measuring key physiological signals taken from the user. My key assumption is that relevance information that derives from the selected sensory channels is correlated with user affective behaviour. However, I do not assume anything about the details of the relationship between users' affective responses and topical relevance; the ground truth is systematically built from the accumulated data, using classification and pattern recognition methods, such as Support Vector Machines and K-Nearest Neighbours.

The employed sensory channels range between facial expressions to physiological signals. The end-goal is to model users' affective responses and predict with reasonable accuracy the topical relevance of viewed results without the need for explicit judgements. The initial findings suggest that the prediction of topical relevance, using the above approach, is feasible and, to a certain extent, implicit feedback models can benefit from incorporating such affective features. Overall, I examine the following research hypotheses:

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

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

6.2 Experimental Methodology

For more details the reader is referred to Section §3.6.1.

6.2.1 Design

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

6.2.2 Participants

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

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

6.2.3 Apparatus

For this experiment I used two desktop computers, equipped with conventional keyboards and mouse. The first computer (server) hosted the retrieval system (Verge Engine) (Vrochidis, King, Makris, Moumtzidou, Mezaris and Kompatsiaris, 2008), the two test collections (document & video), an SQL database for storing the interaction data and eMotion. The second computer (client) provided access to the GUI environment of Verge Engine. It also logged participants' desktop actions, starting, finishing and elapsed times for interactions, and click-throughs, using a custom-made script. The script was executed in the background and stored the above information to the server's database.

Additionally, a camera was installed on the setting: a Live! Cam Optia AF web camera, with a 2.0 megapixels sensor. The camera was used for recording the participants' expressions, as well as apply real-time facial expression analysis. Finally, two unobtrusive wearable devices were used to capture participants' physiological signals: (i) Polar RS800 Heart Rate Monitor[1], and (ii) BodyMedia SenseWear® Pro$_3$ Armband (Andre, Pelletier, Farringdon, Safier, Talbott, Stone, Vyas, Trimble, Wolf, Vishnubhatla, Boehmke, Stivoric and Teller, 2006). All devices and systems performed logging using a common system time.

Test Collection & Search Tasks

For the video indexing TRECVID 2007 test collection was chosen, which contains 200 hours of videos. The collection was built using the Dutch television archive and covers a range of video genres, such as news magazine, science news, news reports, documentaries, educational programming, and other (Smeaton, Over and Kraaij, 2007). For the document indexing TREC 9 (2000) Web Track was chosen, which is a 1.69 billion document subset of the VLC2 collection of 10 gigabyte size (Bailey et al., 2003). By introducing both video and textual information I collected and studied affective data that correspond to different sources of media stimuli, thus developing a media independent feedback mechanism.

The original content of the TREC topics was retained throughout this study, which in terms of delimiting the area of search appears to be effective enough. The basic assumption behind the topic frame is that an information need should be treated as static and accurately defined, providing an objective measure for precision-recall. For each media type category four different TREC topics of varying content and type were selected. The participants had the option for each media type to perform two tasks of their choice. For

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

each topic ten relevant and ten irrelevant results were pre-selected, using as selection criterion the ground truth associated with the collection. However, considering the variability of personal relevance judgements, some relevant/irrelevant items might have not been perceived as such. Therefore, the final decision on the topicality of each item was drawn based on the participants' explicit judgements.

Search Interface

For the completion of the search tasks a customised version of Verge Engine (Vrochidis et al., 2008) was used, which worked on top of TRECVID 2007 and TREC 9 (2000) Web Track test collections. Verge Engine was selected for its versatile interactive environment, which combines several basic retrieval functionalities (visual similarity search module, textual information processing module, etc.) with a user-friendly interface, that can support adequately the submission of search queries and the retrieval of results in the TREC standard format. In the employed version of Verge Engine each result was represented by a link, along with a short summary (in the case of documents) or a thumbnail, and the associated meta-information (in the case of videos). The meta-information consisted of the shot-id, keywords and the duration of the video.

The evaluated version followed a layered architecture approach (similar to the one presented in Section §5.3.3). The first layer was responsible for supporting any interaction occurring during the early stages of the search process (such as query formulation and search execution). Any output generated during that phase was presented in the second layer. From there, the participants could select and preview any of the retrieved items. The content of an item was shown in a separate panel in the foreground, which constitutes the third layer of our system.

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

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 multimedia and online searching. The information obtained from it was used to characterise the participants, but not in subsequent analysis. A Post-Search Questionnaire was also administered at the end of each task, to elicit participants viewpoint on certain aspects of the search process. The questions were divided into four sections that covered the search session, the encountered task, the emotional experience (with respect to the viewed content) and the encountered results.

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

Facial Expression Analysis Software

In this study I employed eMotion, an automatic facial expression recognition system (Valenti et al., 2007). Even though eMotion follows the categorical approach I used MU's data, instead of categorical data, for training the models. MU's are low-level category of features, very similar to Ekman's action-units (AU's). MU's measure the intensity of an emotion indirectly, by tracking the presence and degree of changes in all facial regions associated with it. Moreover, MU's allowed to associate the captured facial expressions with a wider range of affective and cognitive states, which are not accounted during the meta-classification that eMotion applies. For more details the reader is referred to Section §3.6.2.

Physiological Signal Processing

In addition to facial expressions, I recorded a range of peripheral physiological signals using a set of wearable devices (see Section §6.2.3), which gave a more fine-grained imprint of participants' affective states. My motivation for using these instruments was to establish a relation between facial expressions (among other sensory data collected) to users' affective responses, with respect to topical relevance. For more details the reader is referred to Section §3.6.3.

6.2.4 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 every session the participants were given an information sheet, which explained in detail the conditions of the experiment. They were then asked to sign a Consent Form and were informed 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 brief tutorial on the use of the search environment, followed by a calibration of the sensory devices (Polar RS800 & BodyMedia SenseWear ® Pro_3 Armband) and the cameras. To ensure that the participants' faces would be visible to the camera at all times I encouraged them to keep a proper posture, by indicating health and safety measures.

Each participant completed four search tasks in total, two for each media type. In every task they were handed four topics and were asked to proceed with the one they considered more interesting. For each topic the participants were asked to evaluate ten preselected results (either exclusively relevant or irrelevant). The assessment was conducted under the assumption that these were retrieved by the search system for the given topic, while considering the relevance criteria provided by the scenario description. To negate the order effects, the task distribution was counterbalanced by using a Latin Squares design. The participants were asked every time to evaluate as many results as possible (if not all) and were given 10 minutes to complete their task, 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, additionally, administered at the end of each session. The participants were encouraged to ask questions and were, once again, informed about their right to withdraw and have any data gathered on them instantly and permanently destroyed.

6.3 Models

One of the major contributions of this study is the exploration of the role of affective feedback in designing multimedia search systems. The modelling goal is to promote a more natural and meaningful interaction by predicting with reasonable accuracy the topical relevance of videos and online documents. I employed sensory data that derive from facial expressions and other peripheral physiological signals as the only feedback information. From the latter signals I extracted a set of features and used them to perform discriminant analysis using a range of classification methods, such as Support Vector Machines (SVM) and K-Nearest Neighbours (KNN) clustering. I do not assume anything about the relationship between these features (which are considered indicative of participants' affective behaviour) and topical relevance, but rather follow a straightforward classification approach, using the ground truth that is associated with the training data.

6.3.1 Support Vector Machines

I used libSVM[2], an implementation of support vector machines (SVM), to predict the category (two classes: relevant or irrelevant) of documents and videos which were viewed by the participants. This approach utilises an efficient method that can deal with difficult and multi-dimensional classification problems. The models were trained using a radial basis function (RBF) kernel, which, among the basic four SVM kernels (linear, polynomial, radial basis function, sigmoid), was considered as a reasonable first choice. Moreover, the RBF kernel is preferable, since it encounters less numerical difficulties and has a limited number of hyper-parameters.

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

Independently of the kernel function and the parameters, three different categories of SVM models were trained: (i) plain models, using different training and test sets, (ii) models using 5-fold cross-validation, to compensate for the small size of the training/test sets, and (iii) two-layer hierarchical SVM models, with five (5WC) and ten weak (10WC) classifiers. The last category of models used n weak classifiers, each trained on a different subset of the training set. The whole training set was then predicted once, and the output of each weak classifier was used to train the meta-classifier. This hierarchical framework improved the accuracy of classification, in all cases, by a small percentage.

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

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

Table 6.1: Features used to represent participants' affective behaviour (in terms of topical relevance)

6.3.2 K-Nearest Neighbours

The implementation of IB1 from the Memory-Based Learning platform TiMBL (Daelemans and van den Bosch, 2005) was used, which applies non-trivial data structures and speed-up optimisations superior to other implementations (e.g., WEKA[3]) of KNN algorithms. In the training phase TiMBL applies a discount function to each feature based on their Gain Ratio. In the test phase it offers the opportunity to choose from a range of metrics to influence the definition of similarity of neighbours. In order to fine tune the performance of the KNN model all combinations of these similarity metrics were used, namely: (i) weighted overlap, (ii) modified value difference metric, (iii) Jeffrey divergence, and (iv) Levenshtein

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

distance. An iteration over the number of nearest neighbours also took place, with $k = 1$ to 15.

6.4 Data Analysis

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

6.4.1 Features

One of the main objective of this study was to develop a sufficiently rich set of features that would allow to determine whether a user perceives an information item (document, video) as relevant to his/her need. This information was obtained using the instruments described in Sections §3.6.2 and §3.6.3. The original number of features that was acquired summed to 53. However, after applying the feature selection only 29, out of the 53, were eventually kept. By doing so, a subset of relevant features was isolated that allowed for the building of robust learning models while introducing minimum noise. The reduction of the feature set also reduced the dimensionality of the problem and, therefore, the time required to apply the learning algorithms.

The features that were used to model user affective behaviour are summarised in Table 6.1. For clarity they are categorised into two groups: (i) facial expression features, and (ii) peripheral physiological features. Due to the different sample rate of each tool the output of each sensory channel was not synchronised with the rest. However, this did not pose a problem in the analysis, since it was not a necessary step in the training process and as long as there were enough training data for each reading/viewing session. Finally, the analysis was performed on a frame-basis.

Facial expression features: In summary, the user affective behaviour is represented by a selection of features that have directly measured values. Most of these attributes have been associated in the past with important affective and cognitive processes. Several studies by Ekman (1999b) have revealed that certain universal facial expressions, when spontaneously displayed, signal emotions of anger, disgust, fear, happiness, and surprise. Facial expressions are generally regarded as essential aspects of social interactions. The face provides conversational signals, which do not only clarify our current focus of attention (Pantic and Rothkrantz, 2003) but also regulate our interactions with the surrounding environment and the organisms that inhabit it.

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

Galvanic skin response represents the activity of the autonomic nervous system (Boucsein, 1992). Changes in the electrical properties of the skin, due to the activity of the sweat glands, is physically interpreted as conductance. The sweat glands, which are distributed all over the skin, receive input from the sympathetic nervous system, making it a good indicator of the level of emotional arousal due to external sensory or cognitive stimuli. Variations in the skin temperature are mainly the result of localised changes in the blood flow, caused by vascular resistance or arterial blood pressure (Kataoka et al., 1998). Skin temperature, which reflects the autonomic nervous system activity, is also another reliable indicator of the underlying affect state.

6.4.2 Preprocessing

For both categories of sensory information, as well as all media types, the data were randomly separated into two sets (training and test) using an equal number of documents. The resulting sets had approximately the same number of feature vectors. By balancing the training and test sets I prevented over-fitting and, additionally, compensated for the originally uneven size of the data sets. To my knowledge, none of the instruments that was used pre-processes the data. Therefore, I had to scale them before applying any classification method, to avoid having attributes in greater numeric ranges dominating those in smaller numeric ranges. Finally, feature selection was performed using as a selection criterion the information gain of each feature and concluded to a representative

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

Table 6.2: Results for models trained on facial expressions (motion units).

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

Table 6.3: Results for models trained on peripheral physiological signals.

set of features, which was used to train a model with the best feature-wise discriminating ability.

6.5 Results

In this section I present the experimental findings of this study, based on 96 search sessions that were carried out by 24 participants. Out of the many results, I report those that refer to the models and present only the questionnaire data that refer to ethical and usability issues of the proposed system. The performance of all models was measured using the standard IR metrics of accuracy, precision and recall. Accuracy was computed as the fraction of items in the test set for which the models' predictions were correct.

6.5.1 Models

For each classification method I present only the model which achieved the best performance among the rest in its category. The results are shown in Tables 6.2 and 6.3. The McNemar's test was applied to check for statistically significant variation against the baseline. Models marked with (*) got significantly different results, compared to the baseline model, with $p < 0.005$. The baseline represents random choice and, since the class of a result can be either relevant or irrelevant, it is set to 50%. Among all the models, the SVM held the best performance, giving a reasonable, though rather noisy, prediction of topical relevance. The boosting that was applied, using either 5 or 10 weak classifiers, gave a slight increase in the accuracy. The model that held the best accuracy, from those trained on facial expression data, was the SVM with 5 weak classifiers (64.9%), followed by the SVM with 10 weak classifiers (57.9%) for the documents category. From the models trained on peripheral physiological signals the SVM with 5-fold cross-validation held the best accuracy for both video (66.5%) and documents (60.4%).

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

6.5.2 Questionnaires

A 5-point scale Likert scale was used in all questionnaires. Questions that ask for participants' ratings on a unipolar dimension have the positive concept corresponding to the value of 1 (on a scale of 1-5) and the negative concept corresponding to the value of 5. Questions that ask for participants' ratings on a scale of 1-5 represent in the analysis stronger perception with high scores and weaker perception with low scores. When asked about the importance (M = 3.2917, SD = 1.1221) and helpfulness (M = 3.3333, SD = 1.0072) of a search system that integrates affect aware technologies, the participants did not exhibit a major trend in their views.

The same applies for their view in terms of privacy and intrusiveness (M = 3.1250, SD = 1.1156), which is a positive outcome, considering that the success of such a system depends highly on its acceptability. One interesting finding is that, overall, participants' do not perceive as unethical to have their emotional behaviour monitored (M = 2.5000, SD = 1.1034). Perhaps this is one aspect of human-computer interaction that imitates human-human interaction and, therefore, feels more tolerable and natural. Finally, the participants consider an affect-aware search system better than existing search tools that do not integrate similar emotion-detection modules (M = 3.5000, SD = 0.7223).

6.6 Follow-up Study

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

6.7 Discussion

In this chapter I have presented a controlled experimental framework where the participants evaluated the relevance of videos and documents, according to the topic at hand. The purpose of the study was to analyse quantitatively the affective responses of users engaging in the assessment of retrieved results, in an attempt to locate relevant items. The experiment was designed to resemble actual search scenarios. Measurements of facial expressions and key physiological signals were taken, and used with classification techniques (SVM & KNN) to train models capable of discriminating successfully the category (relevant vs irrelevant) of the information items.

One facet of affect recognition is developed here for the first time: classification of users' affective responses from facial expression and physiological data, gathered from many participants. The accumulated evidence supports both hypotheses, namely that participants' affective responses, as determined from the observation of their facial expressions and other peripheral physiological signals, will vary across the relevance of perused information items. It also indicates an associated relationship between participants' affective responses and topical relevance. Among all the models, the SVM held the best accuracy (66.5%), giving reasonably better performance than the baseline. I acknowledge that in some cases the content itself might have induced emotional reactions which were unrelated to topical relevance. However, I believe that this issue can be dealt by introducing a multi-modal framework for affective feedback.

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

6.8 Summary

The work discussed here takes the work presented in Chapter 5 one step further. In this chapter I examined whether affective information can be employed as an implicit source of evidence for topical relevance and, additionally, investigated ways of optimising the affective models' performance. For modelling relevance I extracted a set of features from the acquired signals and applied different classification techniques on a number of sensory channels, such as facial expressions and physiological signals. These channels are regarded as indicative of users' affective states. The results of the evaluation suggest that the prediction of topical relevance, using the above approach, is feasible and, to a certain extent, implicit feedback models can benefit from incorporating such affective features.

Chapter 7

User Study 4: A Study of the Effect of Experimental Conditions

Preamble

In this chapter I investigate the effects of the experimental conditions on the affective model's performance. The models are trained using the same set of sensory data presented in Chapter 6. The size of the effect is determined by comparing the performance of models trained on: (i) sensory data from a semi-controlled experiment, and (ii) data from a controlled experiment.

7.1 Introduction

In this chapter I present a continuation of the work discussed in Chapter 6. The main purpose is to assess the affective models' performance, when trained on data gathered under varying experimental conditions. Throughout this thesis I have introduced three different experimental setups: (i) naturalistic, (ii) semi-controlled, and (iii) controlled. The first type involves the carrying out of typical search tasks, under few restrictions (duration of tasks, availability of search topics). In that respect, it grands the participants a higher degree of freedom regarding the actions they can perform, the search queries they can submit, and the result they can retrieve and browse. In addition, the affective data are collected using hidden recording, which contributes to the creation of a naturalistic setup, such as the one presented in Chapters 4 and 8.

The second type involves the implementation of search tasks with pre-defined topics and pre-selected results. The results must be all assessed in an orderly fashion with respect to their relevance, but the amount of time spent on each result is decided by the participant. The absence of meta-data or summaries makes the assessment of the results' content a necessary step for the completion of the task. Moreover, the collection of affective data is performed using open recording, which introduces the participants to an artificial situation. An example of such an experimental setting is presented in Chapter 6.

Finally, the controlled experimental setup introduces the above limitations and, in addition, imposes a constraint on the time spend for the assessment of each result. This condition keeps the participants highly focused and prevents them from deviating from the task objectives. The current chapter provides an example of a controlled experimental setup. Here the employed sensory channels range between facial expressions to physiological signals. The modelling goal is to determine the topical relevance of information items without the help of explicit judgements. For modelling relevance a set of features is extracted from the sensory data and classified using Support Vector Machines. The size of the experimental conditions' effect is determined by comparing the performance of the second and third category of models: (i) models trained on sensory data from a semi-controlled experiment (see Chapter 6), and (ii) models trained on data from a controlled experiment. The results of the evaluation suggest that, in a controlled experimental setup, the participants are not as expressive and spontaneous, thus giving noisy data with no discriminating characteristics.

Overall, I examine the following research hypothesis:

H$_1$**:** Affective models trained on data gathered under semi-controlled experimental conditions will exhibit better performance compared to models trained on data gathered under controlled experimental conditions.

7.2 Experimental Methodology

For more details the reader is referred to Section §6.2.

7.2.1 Design

This study used a repeated-measures design. There were two independent variables: media type (with two levels: "text" and "video") and topical relevance (with two levels: "relevant" and "irrelevant"). The media type levels were controlled by assigning topics associated with a text or a video collection accordingly. Topical relevance levels were controlled by presenting results that belonged exclusively to one of the two categories, using the ground truth associated with the collection (TREC's qrel) as a selection criterion. The dependent variables were: (i) accuracy, (ii) precision, and (iii) recall.

7.2.2 Participants

Twenty-four healthy participants, all employees at the Institute of Telematics and Informatics (ITI), applied for the study through an organisational-wide ad. The participants were in their majority of Hellenic nationality and had a mixed-educational background (3 with PhD degrees, 10 with MSc degrees, 7 with BSc degrees and 4 other). They were all proficient with the English language (2 native, 17 advanced, and 5 intermediate speakers).

Out of 24, 12 were male and 12 were female, and were between 23-39 years of age (*M* = 27.66, *SD* = 3.73). They had an average of 8 years of online searching experience and all claimed to have been using at least one search service in the past (with the most popular being "Google Video" and "Youtube"). On average, the participants reported carrying out document searches once or twice a day (*M* = 5.33, *SD* = 0.96) and carrying out image or video searches once or twice a week (*M* = 4.79, *SD* = 1.21). 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").

7.2.3 Apparatus

For this experiment I used two desktop computers, equipped with conventional keyboards and mouse. The first computer (server) hosted the retrieval system (Verge Engine)

(Vrochidis et al., 2008), the two test collections (document & video), the SQL database that held the interaction data and eMotion. The second computer (client) provided access to the GUI environment of Verge Engine. It also logged participants' desktop actions, starting, finishing and elapsed times for interactions, and click-throughs, using a custom-made script. The script was executed in the background and stored the above information to the server's database.

Additionally, a camera was installed on the setting: a Live! Cam Optia AF web camera, with a 2.0 megapixels sensor. The camera was used for recording the participants' expressions, as well as apply real-time facial expression analysis. Finally, two unobtrusive wearable devices were used to capture participants' physiological signals: (i) Polar RS800 Heart Rate Monitor[1], and (ii) BodyMedia SenseWear® Pro_3 Armband (Andre et al., 2006). All devices and systems were logging using a common system time.

Test Collection & Search Tasks

For the video indexing I used the TRECVID 2007 test collection, which contains 200 hours of videos. The collection was built using the Dutch television archive and covers a range of video genres, such as news magazine, science news, news reports, documentaries, educational programming, and other (Smeaton et al., 2007). For the document indexing I chose Wikipedia XML Collection 2007 (Denoyer and Gallinari, 2006). The selected text passages were between 250-300 characters long. By introducing both video and textual information I allowed for the collection of affective data for different sources of media stimuli, thus developing a media independent feedback mechanism.

I chose to retain the original content of the TREC topics throughout this study, which in terms of delimiting the area of search appears to be effective enough. The basic assumption behind the topic frame is that an information need should be treated as static and accurately defined, providing an objective measure for precision-recall. For each media type ten different TREC topics of varying content and type were selected. The participants had the option, for each type, to perform two tasks of their preference. For every topic 7 relevant and 7 irrelevant items were pre-selected and randomly distributed in the results list, using as selection criterion the ground truth associated with the collection. However, considering the variability of personal relevance judgements, some relevant/irrelevant items might have not been perceived as such. Therefore, the final decision on the topicality of each item was drawn based on the participants' explicit

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

judgements. A list of all the topics is presented in Table 7.1.

Text Topics
Topic 1: Search for articles about sailing and sailboat specific vocabulary.
Topic 2: Find article about grains that describe their nutritional benefits and how they are cooked, or articles about a dish that uses grains as a main ingredient.
Topic 3: Find information about the climate in Norway in summer.
Topic 4: I want information about weather prediction.
Topic 5: Find tourist attractions in Scotland.
Topic 6: Find information about the features of healthy diet.
Topic 7: Retrieve information about references in the simpsons episodes.
Topic 8: Find paragraphs that describe fuel consumption of diesel engines, not other types of engines.
Topic 9: Find information about salad recipes.
Topic 10: I would like to find out the health risks of food additives.
Video Topics
Topic 1: Find shots of one or more people walking up stairs.
Topic 2: Find shots of a person walking or riding a bicycle.
Topic 3: Find shots of hands at a keyboard typing or using a mouse.
Topic 4: Find shots of a street protest or parade.
Topic 5: Find shots with hills or mountains visible.
Topic 6: Find shots of a street at night.
Topic 7: Find shots with 3 or more people sitting at a table.
Topic 8: Find shots in which a boat moves past.
Topic 9: Find shots of a woman talking toward the camera in an interview - no other people visible.
Topic 10: Find shots of a classroom scene with one or more students.

Table 7.1: A list of the available search tasks

Search Interface

For the completion of the search tasks a customised version of Verge Engine (Vrochidis et al., 2008) was used, which worked on top of TRECVID 2007 and Wikipedia XML 2007 collections. Verge Engine was selected for its versatile interactive environment, which combines several basic retrieval functionalities (visual similarity search module, textual information processing module, etc.) with a user-friendly interface, that can support adequately the submission of search queries and the retrieval of results in the TREC standard format. In the employed version of Verge Engine each result was represented by a link (resullt1, result2, etc.), in the case of text results, or a thumbnail, in the case of video results.

The evaluated version followed a layered architecture approach (similar to the one presented in Section §5.3.3). The first layer was responsible for supporting any interaction occurring during the early stages of the search process (such as query formulation and search execution). Any output generated during that phase was presented in the second layer. From there, the participants could select and preview any of the retrieved items. The content of an item was shown in a separate panel in the foreground, which constitutes the third layer of the system. Upon opening a result (in the third layer), the participants had 30 seconds to read the text or watch the video. The time was sufficient for a person to read through the text, or view the video, while remaining concentrated on his/her task. The time threshold was decided while considering the average number of characters in the text passages, the duration of the videos and the reading time (characters per second), based on previous evaluations. The remaining time (in seconds) was also indicated on the panel.

The main purpose of the layered architecture was to isolate the viewed content from all possible distractions that reside on the desktop screen; therefore, establishing additional ground truth that allowed to relate users' affective responses to the source of stimuli (the perused information items). The time restriction also discouraged the participants from deviating from their search tasks; thus staying focused on their task objectives. This was an important aspect of my experimental methodology, since I was interested in isolating content-particular user emotions. Upon viewing a result, the participants had to evaluate its degree of relevance on a 5-point scale.

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 multimedia and online searching. The information obtained from it was used to characterise the participants, but not in subsequent analysis. A Post-Search Questionnaire was also administered at the end of each task, to elicit participants viewpoint on certain aspects of the search process. The questions were divided into four sections that covered the search session, the encountered task and the emotions experienced during the search process. All of the questions included in the questionnaire were forced-choice type. Finally, an Exit Questionnaire was introduced at the end of the study. The questionnaire gathered information on the topic descriptions and the perceived relevance criteria, as well as participants' views of the importance of affective feedback, with respect to usability and ethical issues.

Facial Expression Analysis Software

For more details the reader is referred to Section §3.6.2.

Physiological Signal Processing

For more details the reader is referred to Section §3.6.3.

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

Each participant completed four search tasks in total, two for each media type. In every task they were handed ten topics and were asked to proceed with the one they considered the most interesting. For each topic the participants were asked to evaluate 14 pre-selected, randomly distributed results (7 relevant and 7 irrelevant), while considering the relevance criteria provided by the scenario description. To negate the order effects the task distribution was counterbalanced by using a Latin Squares design. The participants were asked every time to evaluate as many results as possible and were given 10 minutes to complete their task, during which they were left unattended to work.

At the end of each task the participants were asked to complete a Post-Search Questionnaire. An Exit Questionnaire was, additionally, administered at the end of each session.

The participants were encouraged to ask questions and were, once again, informed about their right to withdraw and have any data gathered on them instantly and permanently destroyed.

7.3 Models

One of the contributions of this study is the investigation of affective feedback, under controlled experimental conditions. The modelling goal is to promote a more natural and meaningful interaction by predicting with reasonable accuracy topical relevance of videos and text passages. I employ sensory data that derive from facial expressions and other peripheral physiological signals as the only feedback information. From the latter signals a set of features is extracted and used to performed discriminant analysis, using Support Vector Machines (SVM). I do not assume anything about the relationship between these features and topical relevance, but rather follow a straightforward classification approach using the ground truth that is associated with the training data.

7.3.1 Support Vector Machines

I used libSVM[2], an implementation of support vector machines (SVM), to predict the category (two classes: relevant or irrelevant) of documents and videos that were viewed by the participants. This approach utilises an efficient method that can deal with a difficult, multi-dimensional classification problem. I trained my models using a radial basis function (RBF) kernel, which, based on previous work (see Chapter 6), 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 SVM model a grid-search of the parameters C (cost) and γ (gamma) was performed using cross-validation, during which exponentially growing sequences of C and γ were tried. However, since performing a full grid-search can be time consuming, a coarse grid was used initially and only after identifying a "good" region" a finer grid search was performed. The end-purpose was to identify the optimal set of (C,γ) so that every classifier trained could achieve the best possible (tuning-wise) accuracy score on in testing data. Ten SVM models were trained, using each time differently shuffled training and test sets. The performance scores presented in the results section (§7.5) are average scores across all ten models.

[2]http://www.csie.ntu.edu.tw/čjlin/libsvm/

7.4 Data Analysis

Out of the 2174 browsing instances that took place in this study, 1096 correspond to text passages and 1078 to videos. Overall, for the facial expressions category, 243992 feature vectors were gathered (141354 for text sessions, 102638 for video sessions), while for the biometrics category a considerably smaller set of 59508 feature vectors was accumulated (34426 for text sessions, 25082 for video sessions). Finally, for the HR data a total of feature vectors was gathered (11219 for text sessions, 7733 for video sessions). However, HR has been excluded from this analysis, since the preliminary tests did not reveal any information gain.

7.4.1 Features

For more details the reader is referred to Section §6.4.1.

7.4.2 Preprocessing

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

7.5 Results

In this section I present the experimental findings of this study, based on 96 search sessions that were carried out by 24 participants. Out of the many results, I report those that refer to the affective models' performance. Their performance was measured using the standard IR metrics of accuracy, precision and recall. Accuracy was computed as the fraction of items in the test set for which the models' predictions were correct.

7.5.1 Models

For each category of sensory data I present the average performance of the models, across ten evaluations. The baseline represents random choice and, since the class of a result can be either relevant or irrelevant, it is set to 50%. As show in Tables 7.2 and 7.3, both categories of models held a similar, rather noisy performance, which is comparable to the baseline. Among all the models, the SVM model trained on the peripheral physiological signals for the media type "videos" held the best performance (54.35%), followed by the model trained on facial expression data (53.17%), for the same category. The models that were trained on the *emotion categories* of facial expression data had lower accuracies, unlike the models trained on the MU's, and were omitted from further discussion. I speculate that this was a result of the meta-classification that eMotion applies, when categorising each captured expression into one of the seven emotion categories. Another potential reason for the low accuracy is the reduction of the dimensionality of the feature space, which is a result of the meta-classification applied.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**
Baseline (Random selection)	50.0	0.0	0.0
Videos			
SVM Training & Test Set	53.17	39.0	50.0
Documents			
SVM Training & Test Set	46.45	50.3	45.62

Table 7.2: Results for models trained on facial expressions (motion units).

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**
Baseline (Random selection)	50.0	0.0	0.0
Videos			
SVM Training & Test Set	54.35	48.12	51.22
Documents			
SVM Training & Test Set	45.92	41.41	43.2

Table 7.3: Results for models trained on peripheral physiological signals.

7.6 Discussion

The accumulated evidence supports my hypothesis, namely that affective models trained on data gathered under semi-controlled experimental conditions will exhibit better performance compared to models trained on data that were gathered under controlled experimental conditions. The performance, in terms of accuracy, of all models was found to be comparable to the baseline (50%), suggesting that the experimental conditions had a direct impact on the quality of the affective data. This suggests that, under such controlled conditions, the participants' capacity to interact on an affective level and display authentic facial expressions was substantially hindered. In respect to the physiological metrics, the time window that was given for each result was perhaps not large enough to allow for capturing the micro-momentary physiological changes that followed each relevance judgment. This resulted in poor and noisy data, with no discriminating characteristics.

Regarding the performance of the models for the video topics, it appears that audio-visual content is a stronger stimuli compared to textual information. The difference in the performance between these two categories of models was found greater than 8%. As discussed in Section §6.7, this indicates that multimedia retrieval might prove a more suitable area of application for affective feedback. It also supports further the view that a naturalistic experimental setting would be in favour of acquiring spontaneous and authentic facial expressions (Arapakis et al., 2008), as well as affective data with more discriminative power.

7.7 Summary

In this chapter I have presented an experimental framework where the participants evaluated the relevance of text passages and videos according to the topic at hand. The purpose of the study was to analyse quantitatively the affective responses of the participants, while they were evaluating the results in an attempt to locate relevant items. The experiment was designed to resemble actual search scenarios. Measurements of facial expressions and key physiological signals were taken under controlled experimental conditions and were used to train models that can discriminate between relevant and irrelevant results.

The evaluation of the experimental conditions on the models' performance provides evidence in favour of the research hypothesis examined in this chapter. The findings suggest that, in a controlled experimental setup, the participants are not as expressive and spontaneous; thus giving noisy data with no discriminating characteristics, which result in affective models with poor, or very close to random, performance.

Chapter 8

User Study 5: General Vs Personalised Affective Models

Preamble

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 chapter I examine ways of personalising affective models trained on facial expression data. Using personalised data I 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 I extract a set of features from the facial expression data and classify them using Support Vector Machines. The experimental results indicate that accounting for individual differences and applying personalisation introduces in most cases a noticeable improvement in the models' performance.

8.1 Introduction

The main challenge 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 (Belkin, 1980). 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 (Koenemann and Belkin, 1996a; Rui and Huang, 2000) to implicit (Agichtein, Brill and Dumais, 2006; Badi et al., 2006) (see Sections §2.4.1 and §2.4.2). 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 (Joachims et al., 2005; Morita and Shinoda, 1994). 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. Oard and Kim (2001) defined a set of application-specific observable behaviours (examination, retention, etc.) and introduced the concept of learning user interests and building user profiles from implicit data. Puolamäki, Salojärvi, Savia, Simola and Kaski (2005) combined 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.

Agichtein, Brill, Dumais and Ragno (2006) 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, Yu and Meng (2002) constructed user profiles based on users' search history and developed algorithms that map query terms to predefined categories. The latter information was used to extract users' interests and address issues related to word ambiguity. Teevan, Dumais and Horvitz (2005) argue 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.

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. Daoud, Tamine-Lechani and Boughanem (2008), considered in this context the problem of search-session boundary recognition. In this 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 chapter I examine different ways of personalising affective models, trained on facial expression data gathered by many individuals. My objective is to determine whether the behavioural differences among users have an impact on the models' ability to discriminate between relevant and irrelevant documents. The work presented here is limited to that of modelling users' affective behaviour and does not involve information filtering or adaptation of content. Using personalised data I adapt these models to individuals and compare their performance to a general model. For modelling relevance a set of features is extracted from the facial expression data and classified using Support Vector Machines. The findings suggest that accounting for individual differences and applying personalisation introduces, in most cases, a noticeable improvement in the models' performance. To my knowledge, no prior work has ever applied personalisation on the affective level interaction, in the context of online information seeking. Overall, I examine the following research hypotheses:

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

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

8.2 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 (Sebe et al., 2007). 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 I devised an experimental setup, similar to the one adopted in Chapter 4, that mitigated most of the unwanted effects. My primary goal was to expose the participants to stimuli of varied intensity. As a result, the information that was collected covered a much wider spectrum of affective behaviour and allowed the comparability of results with previous work. For more details the reader is referred to Section §3.6.1.

8.2.1 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 I 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.

8.2.2 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").

8.2.3 Apparatus

For this experiment I used two desktop computers, equipped with conventional keyboards and mouse. The first computer (server) hosted the retrieval system (Verge Engine) (Vrochidis et al., 2008), the two test collections (document & video), the SQL database that held the interaction data and eMotion. The second computer (client) provided access to the GUI environment of Verge Engine. It also logged participants' desktop actions, starting, finishing and elapsed times for interactions, and click-throughs, using a custom-made script. The script was executed in the background and stored the above information to the server's database. In addition, a camera was installed on the setting: a Live! Cam Optia AF web camera, with a 2.0 megapixels sensor. The camera was used for recording the participants' expressions, as well as apply real-time facial expression analysis.

Search Tasks

A number of search tasks was prepared 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 (Borlund, 2000). By doing so, I introduced short cover stories that helped describe to the participants the source of their information need, the environment of the situation and the problem to be solved. I believe that this way I facilitated a better understanding of the search objective and, in addition, introduced a layer of realism, while preserving well-defined relevance criteria. An indicative list of the topics is presented in Table 8.1.

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 8.1: A list of the available search tasks

Search Interface

For the completion of the search tasks I 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 returns 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 their content.

Zoogle applies a layered architecture approach, similar to that adopted in Chapter 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 the 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 to relate participants' affective responses to the source of stimuli (in this case, the perused documents). This was an important aspect of my experimental methodology, since I was 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.

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. The information obtained from it was used to characterise the participants, but not in subsequent analysis. A Post-Search Questionnaire was 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 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.

Facial Expression Recognition

For more details the reader is referred to Section §3.6.2.

8.2.4 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 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 they were encouraged 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 §8.2.1). In every task they were handed six topics and were asked to proceed with the one they considered the most interesting. For each topic the participants 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 results. 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). The task distribution was counterbalanced using a Latin Squares design, to negate any order effects. 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.

8.3 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, 440557 feature vectors were collected, out of which 224165 are associated to relevant documents and 216392 to irrelevant documents. The main objective was to accumulate a sufficiently rich and balanced set of affective data that would allow to experiment with different personalisation approaches. The analysis was performed on a frame-basis.

8.3.1 Features

From the output of eMotion I concluded to a subset of 12 features that have directly measured values and were used to train the 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) I did not employ categorical data for the training of my models. Instead, I used the MU's data, which is a low-level category of features very similar to Ekman's (1999b) action-units (AU's). MU's measure the intensity of an emotion indirectly, by tracking the presence and degree of changes in all facial regions associated with it. Moreover, MU's allowed to associate the captured facial expressions with a wider range of affective and cognitive states, which are not accounted during the meta-classification that eMotion applies.

8.3.2 Preprocessing

The data of each participant was shuffled and split into three subsets, two of which were used for training purposes (S_1 & S_2) and one for testing (S_3). Each time an equal number of documents was used. The datasets S_1, S_2 and S_3 were also resampled, 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 I prevented over-fitting and, additionally, compensated for the originally uneven size of the datasets. Since eMotion does not pre-processes the data I had to scale them before applying any classification method, to avoid having attributes in greater numeric ranges dominating those in smaller numeric ranges.

8.4 Models

I explore the effect of personalisation on the affective models' performance. The modelling goal is to promote a more natural and meaningful interaction by predicting with reasonable accuracy the topical relevance of online documents. Sensory data that derives from facial expressions is employed as the only implicit feedback information. From the latter signals I extract a set of features, and perform discriminant analysis, using Support Vector Machines (SVM). I do not assume anything about the relationship between these features but follow a straightforward classification approach, using the ground truth that is associated with the training data.

8.4.1 Support Vector Machines

I used libSVM[1], an implementation of SVM, to discriminate between two classes of documents: (i) relevant, and (ii) irrelevant. This approach utilises an efficient method that can deal with a difficult, multi-dimensional classification problem. I trained my models using a radial basis function (RBF) kernel, which, based on previous work (see Chapter 6), proved to be the optimal choice. Moreover, the RBF kernel is preferable, since it encounters less

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

numerical difficulties and has a limited number of hyper-parameters.

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

8.4.2 Personalisation

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

In the first approach a total of 19157 instances of general data was used, which was acquired from the user study presented in Chapter 4, in combination with 4300 instances (in three sets of 1430 instances) per participant. For every participant I 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 I 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, the same process was repeated using N+N instances of general data, or N^*+N^* instances of personalised data. This way I was able to examine if the addition of personalised data improved the performance of the model more, compare to adding general data.

In the second approach I examined whether predictions from two different sources (general model and personalised model) could be fused, on a decision level, to predict the topical relevance of a document. For each participant a personalised SVM model was trained, 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. I 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):

$$p^i_{gen} \cdot w_{gen} + p^i_{pers} \cdot (1 - w_{gen}) = p_i \tag{8.1}$$

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. I then calculate the probability p_i of the instance i being relevant. Where $w_{gen} \in [0, 1]$, is the weight I 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.

8.5 Results

In this section I present the experimental findings of the study, based on 48 search sessions that were carried out by 16 participants. Out of the many results, I am reporting those that refer to my models and present only the questionnaire data that refer to the tasks and the search process. The performance of the models was measured 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.

8.5.1 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 participant rating on a scale of 1-5 represent in the 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.

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 8.2: Descriptive statistics on tasks

Task	Stressful - Relaxing		Interesting - Boring		Tiring - Restful		Easy - Difficult	
	M	SD	M	SD	M	SD	M	SD
T_1	2.1250	0.9574	2.5625	1.4592	2.4375	1.1529	1.9375	0.9979
T_2	2.3125	1.0145	2.4375	1.2633	2.8125	1.1087	2.0000	0.9661
T_3	2.0625	1.1236	2.0625	1.2366	2.6875	1.0782	2.5000	1.0328

Table 8.3: Descriptive statistics on the search process

Tasks

Table 8.2 shows the means and standard deviations for participants' assessment of the search tasks. With respect to task difficulty 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 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 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 regarding the following questions: (i) "Overall, the results that were presented to you were: (Range: 1-5, Lower = Relevant)", and (ii) "You feel satisfied with the retrieved results (Range: 1-5, Lower = Agree) ". 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 8.2 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.

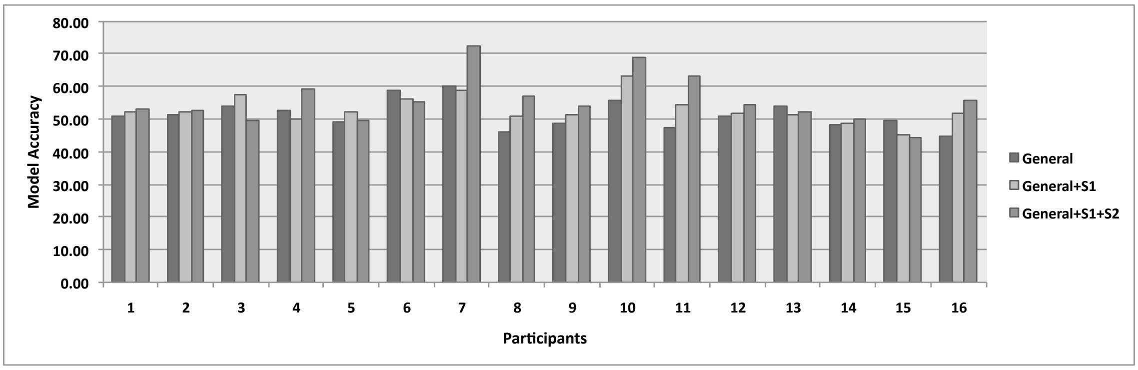

Figure 8.1: Results for models adapted using personalised data

Search Process

Similarly to the task, I examined the effect of the independent variable on the search process. Table 8.3 shows the means and standard deviations of participants' assessment on search process stressfulness, interest, fatigue and difficulty. The Friedman's ANOVA test was applied for the above aspects but did not reveal any statistically significant difference across the tasks, apart from the aspect of difficulty ($\chi^2(3, N = 16) = 7.714$, $p < .05$). The Wilcoxon tests did not indicate a significant difference between any of the tasks.

8.5.2 Models

For each personalisation approach I present the performance of the affective 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.

Adaptation

The results of the first approach are shown in Figures 8.1, 8.2, 8.3, 8.4 and 8.5. Figure 8.1 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 I tested these three combinations against the corresponding S_3. Only in this case, the subsets S_1-S_3 were not balanced. Therefore, the contribution of personalised data by each participant varied. The graph in Figure 8.1 suggests that, in most cases, an improvement was achieved by introducing personalised data to the training set, reaching classification rates that exceed 70%.

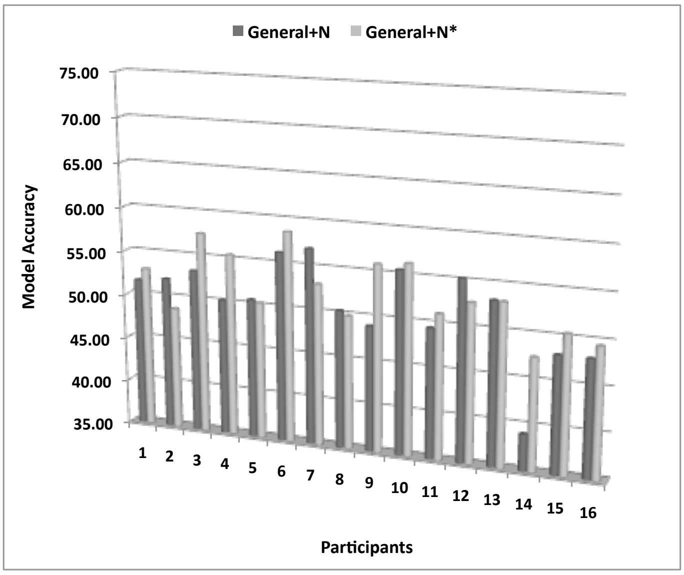

Figure 8.2: Performance of general model after adding N general or N^* personalised data

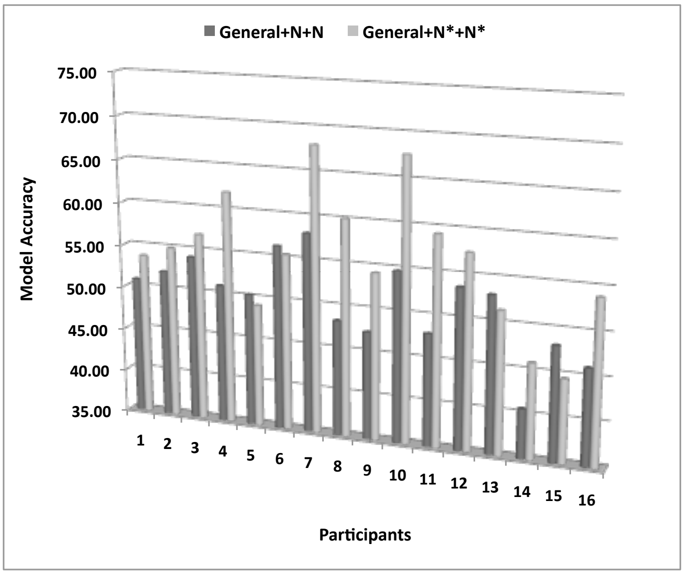

Figure 8.3: Performance of general model after adding N+N general or N^*+N^* personalised data

Figures 8.2 and 8.3 show the results for the same personalisation approach as described above, with the exception that this time balanced sets of personalised data were used (the datasets S_1, S_2 and S_3 were re-sampled 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 8.2 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

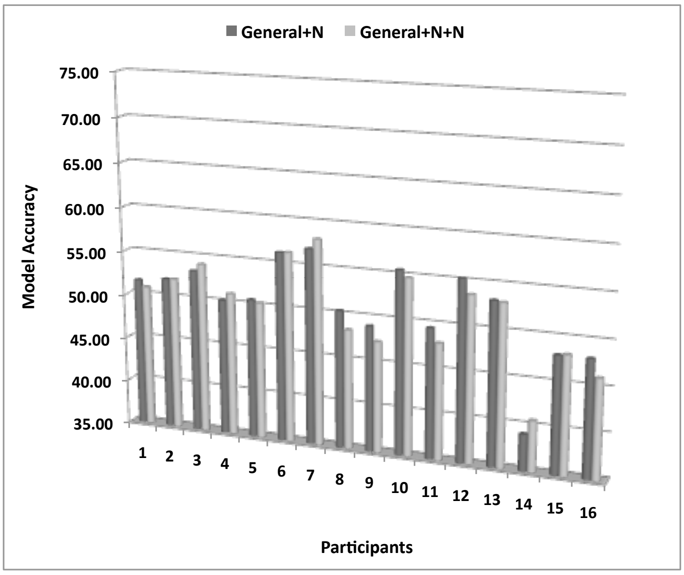

Figure 8.4: Performance of general model after adding N or N+N general data

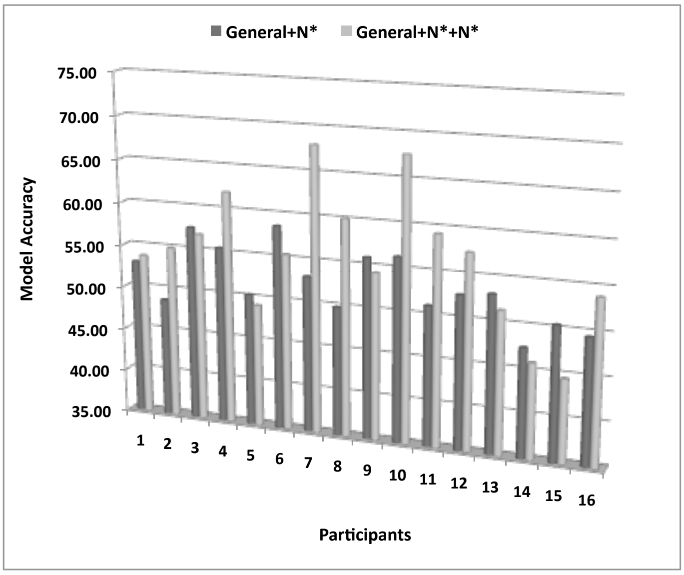

Figure 8.5: Performance of general model after adding N* or N*+N* personalised data

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). This finding indicates that by adding N number of personalised data a slightly better performance was achieved, compared to adding the same number of general data. However, the post-hoc tests did not reveal a significant difference.

Figure 8.3 illustrates the performance of the two classifiers after adding to the original training set N+N general data, or N*+N* of personalised data, and testing against data set S_3, for every participant. The results suggest 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.

Figure 8.4 shows the variation of the classifier's performance after adding to the training set N (M = 51.23, SD = 4.33), or N+N, general data (M = 50.83, SD = 4.34). These two experimental conditions yield a high correlation coefficient (r=.964), which is highly statistically significant ($p < .000$). From this graph it is evident that the classifier's performance did not change after the addition of either N or N+N general data to the original training set.

Figure 8.5 shows the performance of a classifier trained on general data, merged with N^* or N^*+N^* personalised data. On average, the performance of the second classifier (M = 55.94, SD = 6.62) was significantly higher than the performance of the first classifier (M = 52.74, SD = 3.31), $t(15) = -2.147$, $p < .05$. These two examples suggest that the addition of personalised data introduced, in most cases, an improvement in the classifier's performance, as compared to the addition of general data that had no significant effect.

Weighted voting

The results of the second approach are presented in Figure 8.6. In this approach, the general and the personalised data were used separately, to train two different classifiers and combine their predictions using weighted voting. The graph illustrates the performance (y-axis) of the classifiers for different thresholds (x-axis). Each line in the graph is a different weight combination, e.g., the first line is a combination of w_{gen}=0.0 and w_{pers}=1.0. Along the y-axis 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 fourth classifier, with w_{gen}=0.3 (w_{pers}=0.7) and 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.

8.6 Discussion

In this chapter I examined two different approaches to personalising affective models that are capable of discriminating between two categories of documents: relevant and

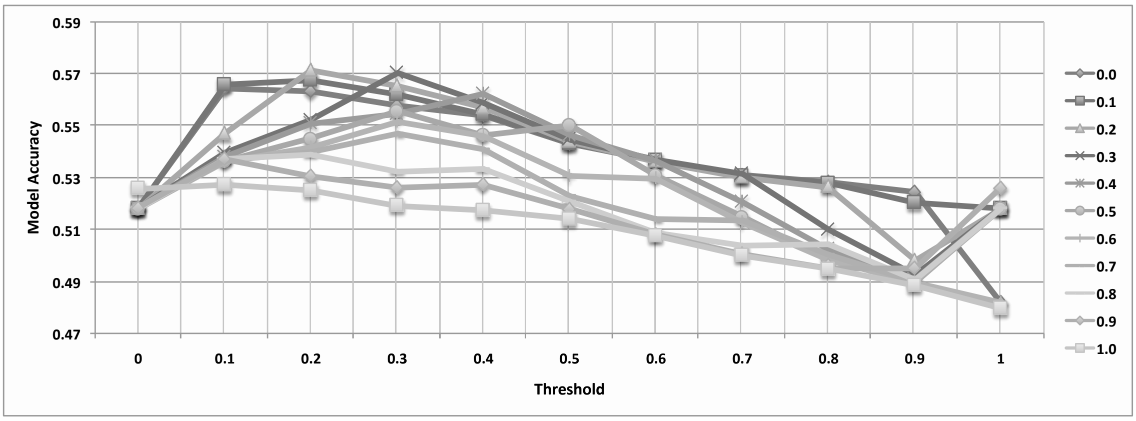

Figure 8.6: Results for the weighted voting (for different combinations of weights and threshold)

irrelevant. For modelling relevance I extracted from facial expression data a set of features and classified them using Support Vector Machines. In the first approach I adapted a general model to the behavioural characteristics of a number of participants, using personalised data, and tested its performance against a model trained exclusively on general data. In the second approach I trained a general and a personalised model separately and combined their predictions using a combination of weighting schemes. Finally, I examined the effect of personalisation on the model's performance and established its 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. The experimental findings support the first hypothesis, namely that by adapting a general affective model to a specific participant a noticeable improvement is introduced in it's discriminating ability. The best performing model attained an accuracy of 72.52%, which is substantially better than the baseline or any other performance presented in Chapter 6. This difference was found to be highly statistically significant, which is an encouraging finding.

Using weighted voting additional evidence was provided in favour of accounting for the behavioural differences of participants. The 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. In respect to the second hypothesis, I cannot suggest which approach was more effective since the 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 participants than others. I speculate that the variation in the models' performance might be correlated with the ability of participants to behave naturally and be expressive in a laboratory setting, similarly to a more familiar environment. However, the choice of setting was necessitated by the need to allow for comparability between data from previous studies.

8.7 Summary

In this chapter I examined two different approaches to personalising affective models that are capable of discriminating between two categories of documents: relevant and irrelevant. I devised an experimental setup that exposed the 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. The post-hoc analysis also indicated that this variation was perceived by the participants, as it was found statistically significant. In conclusion, the work presented here provides some grounds for considering personalisation of affective feedback as a promising area of research and a critical step for the improvement of other facets of the IR process.

Chapter 9

Affective Model of Browsing

Preamble

This chapter presents the Affective Model of Browsing. It is the formalisation of the ideas and the experimental evidence discussed in preceding chapters. The model is of the iterative update of the affective state, resulting from its exposure to, and appraisal of, new information. The model identifies components and processes and describes how these come together in an intuitive way.

9.1 Cognitive Appraisal and Emotional Effects

With the emergence of the cognitive paradigm more attention was brought to the cognitive structures and processes. These structures, highly dynamic and changeable in their nature, are manifestations of human cognition, reflection or ideas and are considered an important aspect of reasoning, decision- and sense-making. The latter processes, which encompass cognitive and social dimensions, occur when the individual deals with new problems, opportunities, or tasks, and involves finding the important structure in a seemingly unstructured situation (Furnas and Russell, 2005).

In the context of IR, such situations involve a human user who, driven by an anomalous state of knowledge, attempts to solve a work task or deal with a perceived problem. The resulting cognitive situation can induce a certain amount of uncertainty and frustration to the user, as to how to reach his goal. This results in a need for external information (information need).

The cognitive approach posed critical research questions and introduced new concepts to the IR community, such as the information need type, the work tasks and the semantic entities, but more importantly it implied the existence of elements and processes other than solely cognitive. Saracevic (1996) suggests than IR interactions occur in several interconnected layers, such as the cognitive, situational and affective. On the cognitive level the information resources are considered as cognitive structures which the user interprets, understands, assimilates and processes cognitively. However, the effects of cognition, especially in information processing tasks, is not limited only to changes in the knowledge structures of the user. On the affective level users interact with intentions, motivations and feelings of urgency, satisfaction, frustration, and other. Immediate emotions can be anticipated after a decision is made, exerting an effect on the mental processes involved in making choice (Kahneman, 2000). Relevance inferences at other levels are often governed by the emotional mechanisms that occur on the affective level.

McKechnie et al. (2007) suggest that the affective variables can play an important role in reading-related information behaviour, especially in the domain of everyday life. Information processing, which occurs during the appraisal process of a goal, an event, or an item, can result in a series of changes in the user's cognitive and affective states. According to Scherer's (2001) component process model (CPM), emotions are elicited and dynamically patterned as the individual appraises continuously and recursively events, situations, or objects, with respect to their effect on his/her values and goals (goal/need conduciveness). Each such check results in the modification of the state of all, or most, of the following organismic subsystems: (i) the cognitive system (appraisal), (ii) the autonomic system (arousal), (iii) the motor system (expression), (iv) the motivational system (action tendencies), and (v) the monitor system (feeling). The emotional reaction is regarded as the cumulative result of such small-scale adaptations of the central, autonomic, and somatic nervous systems, which are linked directly to the results of cognitive appraisal (Aue, Flykt and Scherer, 2007).

Emotions provide the decision maker with evaluative information about the object of appraisal. They are nature's way of directing attention to new information, making new knowledge accessible, and implementing motivations for actions. Once the focus to relevant aspects has been established, further emotions, motivations, or actions, may arise (Pfister and Böhm, 2008). Essentially, emotions constitute a synchronisation of the involved subsystems, driven by the results of cognitive appraisal (Sander, Grandjean and Scherer, 2005). Appraisal can hold over time changing results and changing effects on the subsystems. It is considered the initiator of the synchronisation process (effects on attention, memory, and other cognitive processes) but also the recipient of whatever effects this process produces (modulation of appraisal criteria by emotion components).

I argue that the cognitive, information processing component and the emotion component are not two completely independent, but rather interacting, systems. And although one might regard the brain as a black box, the effects of cognitive and evaluative processes are made visible as manifestations of psycho-physiological changes, accompanied by a series of more or less observable cues (facial expressions, body movements, localised changes in the electrodermal activity, variations in the skin temperature, etc.). The interpretive process can generate a meaning that can potentially affect one's sensory motor automatisms, due to the cognitive content of some emotions (Pfister and Böhm, 2008). This point becomes important in the next section, where I introduce the affective model of browsing.

9.2 The Components of the Affective Model of Browsing

The affective model of browsing provides an approach to capturing an information need that is assumed to be developing during a search session. It relates psycho-physiological changes that occur due to the cognitive appraisal of information encountered in information seeking activities. The model has been built on both theoretical and empirical grounds, most of which have been discussed in previous chapters.

Lets define "affective feedback" as

> *the psycho-somatic mobilisation that occurs during the cognitive appraisal of an information item, with respect to its topical relevance to one's information need.*

Even though the model has been formulated as an analogy to the ostensive model of developing information-needs (Campbell, 2000), in several ways it extends and enriches it with new components and processes. The affective model of browsing is presented diagrammatically with associated propositions and assumptions. The core components are shown in Figure 9.1.

Let A denote an affective-cognitive state, with an emphasis on the affective aspect. I propose this dual component because it encapsulates the inter-dependency between the affective and cognitive elements at play. Emotion should not be treated as merely an external, non-rational force that interrupts an otherwise non-emotional, rational process (Pfister and Böhm, 2008). As discussed in the previous section, the emotional mechanisms are ubiquitous in cognitive appraisal and their functions are multifaceted within every decision making process; thus, the dual character of this component. Within the context of information seeking the following proposition is made:

> **Proposition 1:** Of all the action tendencies present in a user's affective state, the strongest ones are those pertaining to the information need most directly.

Let a denote the actions referred to in P_1. With respect to those actions the following is proposed:

> **Proposition 2:** The actions motivated by an information need are those that will most likely obtain from the environment information that will be regarded by the user as topically relevant.

Let us define the environment, within which these information seeking activities occur, to that of a set of information items. The type of these items can vary between textual (electronic documents, text passages) to audio-visual (images, videos, etc.). Let i denote the information that each item contains. In this environment the user is limited to selecting and viewing the information items. Thus, as P_2 suggests, the selection of an item would indicate its potential relevance to the user's information need.

Let a denote the action of selection. By selecting to view the contents of an information item the user is exposed to new information. However, unlike in the ostensive model, I argue that it is not merely the exposure to new information that results in a new affective state, but rather the appraisal of the item's content with respect to the underlying information need. Let c denote the cognitive appraisal.

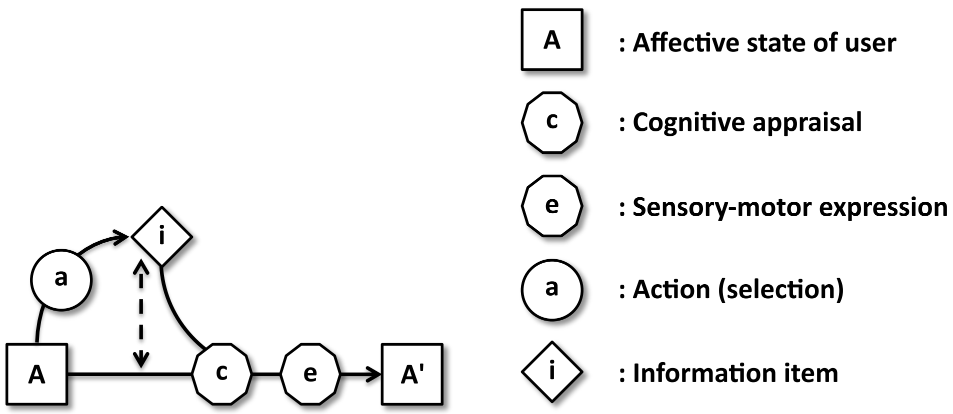

Figure 9.1: The updating of an affective state through the cognitive appraisal of information

The results of the appraisal check will trigger the synchronisation of the subsystems constituting the emotion episode. As the user will reach closer to, or deviate from, satisfying his/her information need, he/she will experience micro-momentary changes in his/her affective state. This psycho-physiological mobilisation is reflected by a series of more or less observable cues, such as facial expressions, body movements, localised changes in the electrodermal activity, variations in the skin temperature, and many more. Let's denote these changes as e. The latter changes will lead to a new affective state A'.

As in the ostensive model (Campbell, 2000), this model does not make any assumptions as to exactly how the processes involved in c and e occur, but rather suggests that the progression from A to A' is due to the latter. With respect to the above the following is proposed:

> **Proposition 3:** Given P_1 and P_2, the majority of changes in A, resulting from the appraisal of i, will be directly related to the information need.

The new affective A' will most likely motivate an action a' involving the selection of a new information item i', which through the cognitive appraisal c' will result in further micro-momentary physiological changes e'' and, thus, to a new affective state A''.

Due to the recursive nature of appraisal the affective and cognitive components are affected by prior changes induced by previous appraisal checks. However, since the significance of an information item can vary from low to high, depending for example by the number of goals or needs that are affected by it, so do the experienced affective changes and the accompanying bodily symptoms will vary in intensity and duration. Therefore, the following is proposed:

> **Proposition 4:** Given P_3, the intensity of the micro-momentary psycho-physiological changes observed in *e* will vary according to the importance of the information need to the user and the degree of relevance of *i*.

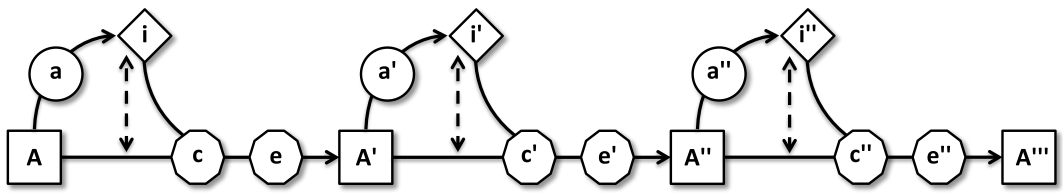

Figure 9.2: The iterative updating of an affective state

9.3 Observable behaviours

Although human emotion behaviour and physiological response patterns are observable there are no objective or direct methods of measuring the subjective experience. The affective state *A* and the cognitive appraisal *c* are two components of the model which are not directly accessible or transparent. However, we can infer the nature and process of these elements by studying user affective behaviour and capturing the, more or less, observable cues discussed in Section §9.1. Considering the latter taxonomy, we can distinguish between two kinds of components: (i) those that are directly observable, and (ii) those that are not directly observable. This distinction is made visible in Figure 9.3 with the use of a dotted line, which separates those components that are internal to the user.

The action *a* of selecting an information item *i* for evaluation, from a limited set of results, can be easily observed. Additional information can be gathered, in terms of which and how many items were selected/rejected. The content of *i* can be also analysed, even though the extracted information may vary depending on the format (textual, image, or video). Furthermore, the affective information communicated during *e* can be also captured through the application of a multi-modal analysis, using a range of sensory channels and devices (see Section §3.4).

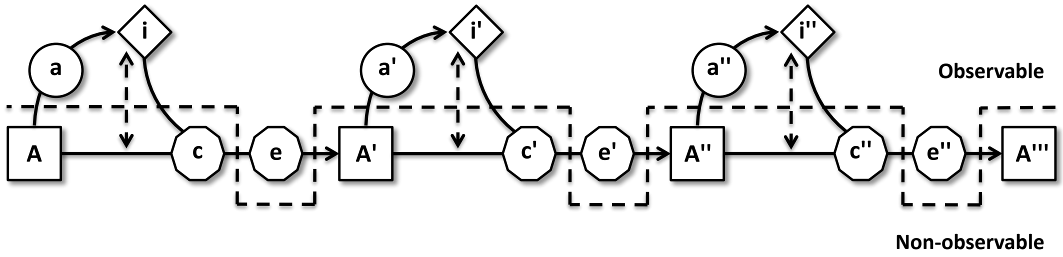

Figure 9.3: Observable and non-observable components

The remaining components, i.e., the affective state *A* and the cognitive appraisal *c* belong to the class of elements that are not directly observable. Nevertheless, we can acquire glimpses at certain qualitative aspects by examining their manifestations, namely the *a* and *e*. P_1 states that the strongest action tendencies *a* are those pertaining to the information need at hand. This suggests that *a* is indicative, at some level, of users' intentions and perhaps some of the aspects of *A*. However, due to the inherent difference in their character (*a* is much more cognitive in nature, while *A* is emphasising the affective aspects of human behaviour) I argue that *a* should not be regarded as much indicative of *A*, as *e*.

The presence of *e*, unlike in the ostensive model, suggests the existence of a range of sensory channels which may convey potentially valuable information regarding users' action tendencies, views, intentions, motivations, feelings, or other. Therefore, the behaviour analysis is not limited to that of interaction data, which as argued in Sections §2.4.1 and §2.4.2 introduces several cognitive-, validity- and robustness-related issues. Moreover, any behaviours captured at this level of interaction should be regarded as qualitatively richer than any other interaction data, due to their spontaneous and involuntary character, as well as their direct relation to the outcome of the appraisal process (see Figure 2.4).

Finally, according to P_2, the selected information *i* will most likely be regarded as relevant to the information need. Thus, *i* is also indicative of the information need. Also, given P_2, the progression from *A* to *A′*, resulting from the appraisal of *i*, will affect the information need most. This suggests that *i*, not only being indicative of *A*, will also be definitive of *A′* along with *e*. We have now *e*, *a* and *i* being indicative of their associated *A*.

However, it would be unrealistic to expect extracting a large amount of unambiguous information from instances of *a*, *i* and *e*. Each of these components introduces an amount of uncertainty and noise, and the possible interpretations of (*a*, *i*, *e*) can be many. After several iterations of *A* to *A′*, several instances of the above components will exist and be available for observations. When studied individually, each triplet (*a, i, e*) will carry with them the same ambiguity, but when taken together the triplets will exhibit ambiguity resolution characteristics. In a similar sense, when examining several information items indicated as relevant, one can infer more accurately the underlying interest compared to what would be possible with only one available item.

9.4 Uncertainty in observable evidence

The proposed model suggests that information needs can be disambiguated and defined through the accumulation of observational evidence. The evidence varies from interaction to sensory data (affective feedback) and can be gathered without the intentional participation of the user; without imposing the cognitive load of explicit relevance judgments; and without exposing the user to internal representation methods used by the IR system. It would, therefore, allow the user to concentrate solely on the task of identifying relevant information.

Since users are better at identifying something as relevant, rather than describing their information needs, it would require several cycles before the system can adapt to their needs. However, this process won't involve the intrusive extraction of descriptions or the explicit indications or information items as relevant or irrelevant. Users' affective responses are regarded as the primary source of implicit feedback and the main indicator of topical relevance. It is, therefore, intended to replace the communicative and explicit actions with observational evidence, deriving from user behaviour.

Let's assume the three iterations presented in Figure 9.2, each of which is represented by a triplet: t_1 = (*a, i, e*), t_2 = (*a′, i′, e′*), and t_3 = (*a″, i″, e″*). The iterations take place in an orderly fashion: first t_1, then t_2, and finally t_3. The time ordering of the evidence suggests the relative degree of uncertainty attached to each instance of triplet, even though it is impossible to determine it absolutely.

After the first iteration, the current affective state is *A′*. *a*, *i* and *e* are the only available evidence at this point that can allow us to make inferences about *A′*. After the second iteration we will be trying to make inferences about *A″*, based on the accumulated evidence (*a, i, e*) and (*a′, i′, e′*). Therefore, as more affective evidence is accumulated, the uncertainty of having an accurate representation of the information need decreases.

As in the ostensive model (Campbell, 2000), there are no solid grounds to suggest that the degrees of uncertainty associated with inferences from t_1 are greater or less than those from t_2 or t_3. The triplet (*a, i, e*) says as much about *A′*, as (*a′, i′, e′*) about *A″*. The model rather shows the path of all the collected evidence leading to the current affective state *A″*. Considering there is a more direct path connecting (*a′, i′, e′*) to *A″*, than there is from (*a, i, e*), it is reasonable to assume that the uncertainty associated with inferences about *A′* based on the latter triplet will be higher than the uncertainty associated to inferences drawn from the former triplet. Furthermore, as the current affective state changes from *A* to *A′*, *A* becomes of little importance. All this can be summarised in the following proposition:

> **Proposition 5:** As the age of affective evidence increases, the uncertainty attached to inferences drawn from it about the current affective state will also increase.

P_5 suggests that the uncertainty attached to evidence increases with the age of that evidence.

9.5 Summary

In this chapter I presented an iterative model of affective state update, based on the cognitive appraisal of viewed information and the progression of the search process. The role of cognition and the effect of emotional mechanisms on mental processes and information processing tasks were also highlighted. The components of the model were characterised as either observable or non-observable. The distinction between the Affective Model and traditional RF is made evident . Finally, the uncertainty associated to inferences drawn from the affective evidence is discussed.

Chapter 10

Ethical & Social Dimensions of Emotion Recognition

Preamble

This chapter addresses briefly some of the ethical and privacy issues that arise from the social-emotional interaction between human users and computer systems. The theme of the discussion is based on: (i) findings that people interact socially and emotionally with machines, even though the latter may have not been designed for such purposes, and (ii) technological advances which allow computers to sense, recognise and respond to emotional and social stimuli. Sensing affect raises critical concerns with respect to social and ethical acceptability, which need to be taken into consideration when designing affect-aware technologies.

10.1 Ethical & Social Acceptability

Traditionally computers have been viewed as tools that facilitate the performance of various information tasks. Since users perceive them as mere machines they have not experienced any social or emotional concerns when interacting with them (Picard et al., 2002). However, a number of studies has shown that users do not respond to computers as tools, but rather employ a wide range of social norms, attitudes and emotional behaviours that guide their interactions (Reeves and Nass, 1996; Picard, 2001; Picard et al., 2002; Vinciarelli et al., 2009). Interfaces can induce a wide spectrum of emotional reactions, despite the fact that they might have not been intentionally designed to invoke any. This finding illuminates the socio-emotional aspects of human-computer interaction, even though users might not be aware of exhibiting any kind of emotional reactions, or are denying doing so.

Current affect-aware systems employ sophisticated sensors that are capable of taking an input signal and processing it for some evidence of emotions. The affective information communicated during human-computer interaction can be captured through the application of a multi-modal analysis, using a range of sensory channels and devices. The sensory input can vary from facial expressions to localised changes in the electrodermal activity. However, computer systems lack the ability to perform moral judgments or act ethical on their own. This suggests, that whatever actions they take will be dependent upon the designer's choices, which may be regarded as unethical (Reynolds and Picard, 2004).

Introducing affect-sensing characteristics to interface design raises critical concerns that need to be examined with respect to social, ethical and privacy issues. Let's consider the perspective proposed by Reynolds and Picard (2004), namely that affect-aware technologies encapsulate the developer's ethical and moral decisions in terms of: (i) which emotions are recognised, (ii) who can access the recognition results, and (iii) for what purposes these are used. Such decisions evoke the problem of acceptability of user behaviour capture and analysis, which not only encompasses the problem of moral acceptability but also the acceptability from the point of view of society and users.

The social applications of these technologies raise the issue of social acceptability, which reflects the legal and ethical aspects that are associated to it. It concerns at the least how users feel about such technologies and whether they consider them respectful of their privacy. Therefore, emphasis should be put on finding out what users consider acceptable or not, what should be done, as well as what can be done. The following section presents an analysis of questionnaire results gathered over three user studies, dedicated to the acceptability of affect-aware systems.

10.2 Analysis of Questionnaire Data from Three User Studies

To understand better users' perception of affect-aware technologies and sensors, in light of the social, ethical and privacy issues discussed earlier, I conducted an evaluation that examined a series of effects. The evaluation was performed using questionnaire data gathered over three user studies (Chapters §4, §5, and §6) that have been partly presented in Chapter §6. The users were introduced to typical IR scenarios, which involved different sensors and invoked a wide range of emotions. Printed and online questionnaires were preferred over other methods of enquiry, since they allow to quickly expose a large group of people to a number of hypothetical situations and ideas.

In the broadest sense I wanted to understand more about the ethical acceptability of affect-aware technologies, such as the ones introduced in my user studies. I considered privacy, invasiveness, effectiveness and desirability as factors that might indicate the ethical/social acceptability. I make the assumption that participants will report a greater sense of privacy invasiveness if they consider affect-based interaction as unethical.

10.2.1 Participants

Seventy-four healthy participants of mixed ethnicity and educational background (17 with PhD degrees, 23 MSc degrees, 29 with BSc degrees and 5 with other degrees) enrolled for the three user studies, through an organisational- or campus- wide ad. They were all proficient with the English language (15 native, 45 advanced, 12 intermediate and 2 beginner speakers). Out of the 74, 39 were male and 35 were female,s and were between 18-45 years of age (*M*=27.51, *SD*=5.59). All participants were free from any obvious physical or sensory impairment. They had an average of 8.28 (*SD*=3.09) years online searching experience and claimed, in their majority, to have been using at least one popular (among many) search service in the past.

10.2.2 Questionnaires

A 5-point scale Likert scale was used in all questionnaires. Questions that ask for user rating on a unipolar dimension have the positive concept corresponding to the value of 5 (on a scale of 1-5) and the negative concept corresponding to the value of 1. Questions that ask for user rating on a scale of 1-5 represent in the analysis stronger perception with high scores and weaker perception with low scores.

The participants were given an example scenario of affect-based interaction and were asked the following questions:

> **1.** *Imagine that you could have used a system with affect-aware capabilities, instead of the one you used in this experiment. This system would unobtrusively determine whether a document is relevant to your search, based on a real-time analysis of your affective behaviour (e.g., facial expressions) while viewing its content. To what extent do you think that your search would potentially improve, considering that this system can automatically retrieve documents similar to the ones which you regarded (or felt) as relevant?*
>
> *Given the system described in question "**1**"...*
> **2.** *...to what extent do you think that it would be important to use it for online searching?*
>
> **3.** *...to what extent do you think that it would be helpful to use it for online searching?*
>
> **4.** *...to what extent do you think that it would be unethical to have your affective behaviour monitored during a search?*
>
> **5.** *...to what extent do you think that it would be intrusive, in terms of privacy?*
>
> **6.** *...to what extent do you think that it would be better, compared to existing search services that do not apply such emotion detection technologies?*

With respect to the first question, the reported answers suggest that participants believe that their online searching can benefit from such affect-aware technologies (N=73, M=3.5068, SD=1.0944). When asked about the importance (N=47, M=3.0425, SD=1.122) and helpfulness (N=47, M=3.3404, SD=1.1282) of a search system, such as the one described in the first question, the participants exhibited a slight positive trend in their views.

The opposite stands for their view of the system in terms of privacy and intrusiveness (N=73, M=3.2876, SD=1.1841) (Figure 10.1). One interesting finding is that, overall, the participants do not perceive as unethical to have their emotional behaviour monitored (N=73, M=2.6575, SD=1.1692) (Figure 10.2). Perhaps this is one aspect of human-computer interaction that imitates human-human interaction and, therefore, it is considered as more natural and acceptable. Finally, the participants reported the affect-aware search system as better than existing search tools (N=72, M=3.2777, SD=0.8429).

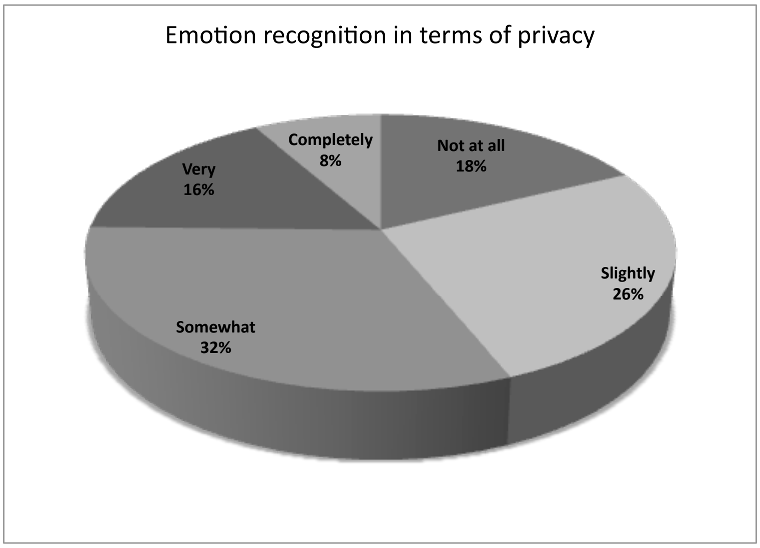

Figure 10.1: Users' perception of how intrusive, in terms of privacy, is to monitor emotions during the search process

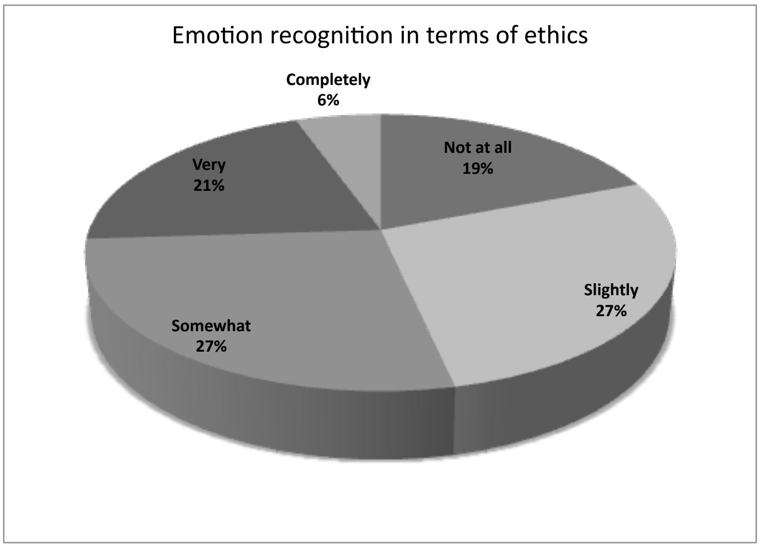

Figure 10.2: Users' perception of how unethical is to monitor emotions during the search process

10.3 Summary

In this chapter I presented the preliminary findings of a series of user studies, which, among other aspects of the search process, addressed the ethical and privacy implications of affect sensors. Overall, the findings suggest that users are reluctant to embrace a viewpoint with respect to whether human-computer interaction should take a more human-like direction. I argue that the acceptability of affect-aware technologies is inhibited by the limited, or non-existing, commercial availability of such applications. This is a result of the technologically profound difficulty in developing systems that can recognise and respond to human emotional communication. Nevertheless, the accumulated evidences reflect a positive, although subtle, trend in favour of the latter.

Chapter 11

Conclusions

Preamble

In this thesis I have investigated the use of affective feedback to aid users interact with search systems more effectively and deal with information uncertainty. The components introduced help users locate relevant results and assist them through meaningful recommendations of unseen items. This chapter summarises the achievements of the work presented in this thesis. It presents the individual achievements made and proposes ways in which the work can be taken further.

11.1 Introduction

This thesis is an exploration of a new area, that of affect-based IR. It is a new field without much pre-existing work or knowledge available.The goal is to aid the information seeking process and improve the experience of a person searching, by investigating the affective aspects of this process. With the work presented here I contribute to the exploration of the role of emotions in the search process and, also, introduce a new approach to the detection and quantification of affective information. This study provides the next step towards the modelling of affective feedback. The following sections gather the achievements reported throughout the thesis and discuss them in detail. These can be summed up to the following:

- Presentation of findings on the role of emotions in the information seeking process.
- Gathering of initial evidence regarding the interaction effect between users' affective responses and task difficulty.
- Validation of the association between users' affective responses and topical relevance.
- Modelling of user affective behaviour for the prediction of topical relevance, in the context of multimedia retrieval.
- Exploration of various sensory channels and classification techniques.
- Investigation of the implications of the experimental setup on users' affective behaviour.
- Comparison of personalised vs general affective models.
- Development of an operational, affect-based, search environment that provides meaningful recommendations based on real-time facial expression analysis.
- Comparison of affective feedback against existing implicit feedback techniques.
- An affective model of browsing that depicts the effects of cognitive appraisal in the information seeking process.
- Investigation into the ethical and privacy implications of affect-based IR.

11.2 User Feedback

In Chapters 2 and 3 I introduced the experimental and theoretical grounds that establish the expression of emotions and the application of social norms as an inherent process of interaction with computing systems and information items. Definitions and theories, as well as methods for studying emotions were also discussed. Additionally, I presented findings from a number of studies, from the fields of HCI, LIS and IR, which indicate that the information seeking process is an integration of three dimensions of the human experience, namely the: (i) affective, (ii) cognitive, and (iii) physical. The same studies suggest a strong correlation of the above dimensions and, moreover, emphasise the regulatory character of the affective component of search behaviour on cognitive processing. Positive and negative emotions were shown to be associated with various aspects of the information seeking process, such as satisfactory search results, successful search completion, content design and aesthetics, and other.

In Chapter 4 I examined the role of emotions in the context of information seeking. In this exploratory study the participants were exposed to realistic search scenarios of varying difficulty, in which they had to evaluate the relevance of online documents according to the topic at hand. The manipulation of task difficulty resulted in a much wider spectrum of affective responses, making the accumulated evidence more authentic and comparable to data gathered in subsequent studies. Questionnaire data were gathered to examine different aspects of the search process, as well as participants' experienced emotional episodes.

Foremost, the study's findings confirm the occurrence of emotions in the information seeking process and suggest that the latter interweave with different psycho-physiological and cognitive processes. In addition, they indicate distinctive patterns that follow a progressive transition from positive to negative valence, as the degree of task difficulty increases. These patterns might prove to be good predictors of significant events and breakdowns that are correlated with changes in the user's knowledge state and information needs. However, further analysis must be performed in order to validate the clarity of this argument.

An analysis of the questionnaire data, along with empirical insights drawn from the study of hidden recordings, suggest that emotions facilitate verbal and non-verbal forms of communication. The context of this communication is indicative of the current state of readiness, attention and active interest towards the search task. Moreover, it reveals the perceived importance of an information item through a psycho-physiological mobilisation, which is reflected by a series of more or less observable, or even unconscious, changes. These changes follow cognitive appraisal and precede any actions taken by the user. On a higher level of abstraction, we could theorise that emotions act as innate mechanisms that help orchestrate users' responses and actions, and shift the focus of attention during breakdowns in the search process.

In the context of RF, it is reasonable to argue that emotions can be invoked by various stimuli, not necessarily associated to topical relevance, with the most prominent being the content of the viewed items. This is an argument which I attempted to address by a careful selection of sensory channels, as well as by studying the interaction effect between the user, the search process and the viewed content. In Chapter 5, I examined the similarities of users' affective responses across different types of stimuli: (i) the search process, and (ii) the viewed content. However, the analysis did not reveal an associated relationship between the reported emotions and the emotional stimuli. Therefore, the experimental evidence is not enough to refute the null hypothesis, namely that user emotions are not consistent across different types of stimuli. This, however, does not necessarily mean that the null hypothesis is true. It only suggests that there is not sufficient evidence against it. In the next section I discuss the user models that were trained on affective data.

11.3 Affective Models

In chapter 6 I investigated the application of affective feedback as an implicit source of evidence for topical relevance. A novel experimental framework was presented that allowed for the quantitative analysis of users' affective responses. My key assumption was that relevance information that derives from selected sensory channels is correlated to affective behaviour. The employed sensory channels ranged between facial expressions to key physiological signals, which are all regarded as indicative of users' affective states. For modelling relevance I extracted a set of features from the acquired signals and applied different classification techniques, such as Support Vector Machines and K-Nearest Neighbours.

One facet of affect recognition is developed here for the first time: the classification of user affective responses from facial expression and physiological data, gathered from many participants. The features that were used to model user affective behaviour are listed in Table 6.1. Both categories of sensory data (facial expressions and physiological signals) proved to be a good first choice, but alternative modalities should be also explored. The multi-modal approach is generally considered superior to uni- or bi-modal approaches, since the sensory channels can complement each other and improve the robustness of the system. With respect to facial expressions, I found that the low-level features perform better compared to high-level information, such as emotion categories.

One potential explanation is that there is unnecessary redundancy of the values produced by the meta-classifier of the automatic facial expression analysis tool, which cannot be modelled efficiently. In addition, a naturalistic experimental setting that applies hidden recording would be in favour of acquiring spontaneous and authentic facial expressions (Arapakis et al., 2008), thus producing data with more discriminative characteristics.

In the evaluation I assessed the performance of the models in determining accurately the topical relevance of viewed items, without the aid of explicit relevance judgements. Among all the models the SVM held the best performance (66.5%), giving a reasonable, though rather noisy, prediction of topical relevance. The boosting that was applied, using either 5 or 10 weak classifiers, gave a slight increase in the accuracy. The model that held the best accuracy (from those trained on facial expression data), was the SVM with 5 weak classifiers, for the documents category. From the models trained on physiological data, the SVM with 5-fold cross-validation held the best accuracy for both video and documents.

The performance of the models for the video topics indicated that audio-visual content is a stronger stimuli, compared to textual information that induced milder affective responses. This suggests that the application of affective feedback in multimedia retrieval might prove more fruitful. The model-based approach presented in Chapter 6 was designed to be as independent as possible from the viewed content and context, therefore, making its application generalizable to a range of different search topics and multimedia.s

Overall, the results of the evaluation suggest that the prediction of topical relevance, using the above approach, is feasible and to a certain extent implicit feedback models can benefit by incorporating such affective features. Furthermore, the integration of affective features could facilitate a more natural and meaningful interaction, improve the quality of the query suggestions, as well as influence other facets of the information seeking process, such as indexing (Chan and Jones, 2005; Hanjalic and Xu, 2005), ranking (Soleymani et al., 2008) and recommendation techniques (Arapakis et al., 2009). Relevance inferences obtained from affective models could potentially provide a more robust and personalised form of feedback. Finally, since there are no other systems available for direct comparison, the system present here holds the best accuracy achieved, so far, in the deduction of topical relevance using affective information.

11.4 An Operational System

A novel video retrieval system, which accounts for user feedback deriving from real-time facial expression analysis, was used to test the performance of affective feedback against existing implicit feedback techniques (see Chapter 5). The search system is realistically applicable; it was implemented using an inexpensive web camera and a standard browser that was modified to communicate with a facial expression recognition system. In addition, it offers a way to process the affective data, using a two-layer hierarchical SVM classifier to discriminate between relevant and irrelevant videos. This information was used to enrich the user profiles and, eventually, generate meaningful recommendations.

The value of the proposed system lies in the combination of different modules and modalities, as well as the seamless integration of affective elements into user profiling. Furthermore, it applies a layered architecture approach that consist of layers dedicated to support different facets of the search process. The main benefit of using this architecture is that it isolates the viewed content from all possible distractions that reside on the desktop screen. It, therefore, establishes reasonable ground truth that allows to associate the captured affective feedback to the source of stimuli. This architecture was applied in all user studies presented in this thesis.

The affect-based retrieval system I propose is an effective way of facilitating and sustaining a different form of RF; one that accounts for the affective dimension of human-computer interaction. The video retrieval system was designed to account for the predictions of the classifiers presented in Chapter 6. The findings suggest that affective feedback, as determined from facial expression analysis, can significantly improve the performance of a recommender system over other popular feedback techniques. The participants favoured more the recommendations offered by the version of the system that accounts for affective feedback, instead of those generated by the baseline that used click-throughs. This suggests a correlation between the facial expressions exhibited by the participants and the topical relevance of the viewed results. This is initial evidence that affective feedback can effectively complement, if not outperform, existing implicit feedback techniques.

11.5 Personalised Models

In Chapter 8 I examined two different approaches to personalising affective models that are capable of discriminating between two categories of documents: relevant and irrelevant. In this experiment the participants were exposed to search tasks of varying difficulty, which was achieved through the re-ranking of the return results. For modelling relevance I extracted from facial expression data a set of features and classified them using Support Vector Machines, similarly to the modelling approach discussed in Chapters 5, 6 and 7.

In the first approach I adapted a general model to the behavioural characteristics of a number of participants, using personalised data, and tested its performance against a model trained exclusively on general data. In the second approach I trained a general and a personalised model separately and fused their predictions using a combination of weighting schemes. Finally, I examined the effect of personalisation on the models' performance and established its significance. The experimental findings indicate that by adapting a general affective model to a specific user we can introduce a noticeable improvement in it's discriminating ability. The best performing model achieved a substantially better performance than the baseline, or any other performance presented in Chapter 6. This difference was found to be highly statistically significant, which is an encouraging finding.

In the second approach I applied weighted voting. The analysis indicated 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. However, no conclusion can be drawn as to which approach was more effective, since the results do 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 suggest that personalisation works better for some users than others. I 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 a more familiar environment. However, the choice of setting was a necessity, guided by the need to allow for comparability of results from previous studies.

11.6 Naturalistic vs Controlled Experimental Setup

In Chapter 7 I presented a continuation of the work discussed in Chapter 6. Essentially, the purpose of this study is to test the affective models' performance when trained on data gathered under varying experimental conditions. The employed sensory channels range between facial expressions to physiological metrics. For modelling relevance I extracted the same set of features as in Chapter 6 and performed classification using Support Vector Machines. The size of the experimental conditions' effect is determined by comparing the performance of two categories of models: (i) models trained on sensory data from a semi-controlled experiment, and (ii) models trained on data from a controlled experiment.

The findings suggest that, in a controlled experimental setup, the participants are not as expressive and spontaneous; thus giving noisy data with no discriminating characteristics, which result in affective models with poor, or very close to random, performance. As discussed in Section §11.3, a potential reason for the noisy performance of the classifiers trained on facial expressions is the overlap that exists in the data. This is a result of capturing subtle to mild emotions that have not reached their peak, due to the controlled experimental conditions. Such emotions naturally trigger facial expressions that are harder to discriminate and, therefore, interpret successfully. A similar phenomenon was also observed in the physiological data. Metrics such as galvanic skin response and body temperature rarely exhibit rapid and intense fluctuations, even less in the case of mild stimuli. Therefore, the effects require a substantial amount of time before they occur. I argue that the overlap in this case was a result of the small time window that was used to acquire the sensory input. In conclusion, a naturalistic experimental setting that applies hidden recording would be in favour of acquiring spontaneous and authentic affective reactions, thus producing data with more discriminating power.

11.7 Ethical Implications

In Chapter 10 I addressed briefly some of the ethical and privacy issues that arise from the social-emotional interaction between human users and computer systems. The theme of the discussion was based on: (i) findings that people interact socially and emotionally with machines, even though the latter may have not been designed for such purposes, and (ii) technological advances which allow computers to sense, recognise and respond to emotional and social stimuli. Sensing affect raises critical concerns with respect to social and ethical acceptability, which need to be taken into consideration when designing affect-aware systems.

I presented the preliminary findings of a series of user studies, which, among other aspects of the search process, addressed the ethical and privacy implications of affect sensors. Overall, the findings suggest that users are reluctant to embrace a viewpoint with respect to whether human-computer interaction should take a more human-like direction. I argue that the acceptability of affect-aware technologies is inhibited by the limited, or non-existing, commercial availability of such applications. This is a result of the technological obstacles in developing systems that can accurately and robustly recognise and respond to human emotional communication. Nevertheless, the accumulated evidences reflect a positive, although subtle, trend in favour of the latter.

The participants, in their majority, expressed the belief that affect sensors can be important and beneficial to their information seeking activities and did not pose any serious ethical concerns. However, their views with respect to privacy issues were slightly different. Participants were concerned with respect to which emotions are recognised, who has access to the recognition results and, finally, the purposes these are used. Emotions, even more than thoughts, are personal and private, revealing information about our most innate motivations and reactions (Picard, 2003). There is an associated a risk of potential misuse of this information to manipulate users, which could constitute a severe breach of ethics. On the other hand, people routinely recognise and respond to emotions, or manipulate them in ways that are regarded as ethical and desirable. Such decisions naturally evoke the problem of acceptability of behavioural analysis, and despite this topic having received much criticism it is argued that interface design should account for the socio-emotional dynamics of interaction (Reeves and Nass, 1996; Picard, 1997; Picard, 2001). Whether interfaces should become more human-like, or just sense and respond to human emotion, remains a subject of debate.

Technology is evolving and being shaped by theories which suggest that human-computer interaction in inherently social and emotional. These views compel more designers to consider the use of affect-sensors in the development of interactive systems. However, as discussed in Section §10.1, this suggests that designers ought to be more thoughtful about social and emotional norms, seek ways to secure users' privacy, and promote less intrusive and natural means of affect-based interaction.

A next step would be to investigate how users feel about affect-sensors when actually using them, instead of being asked about them. According to Reynolds and Picard (2004), future evaluations should focus on uncovering the elements of communication that users consider as determinants of ethical interaction. There are many dimensions and aspects of interaction which can be fragments of what users perceive as ethical and these demand additional exploration.

11.8 Future Work

This thesis has examined a number of issues in the areas of affective feedback and interactive IR. Many avenues have emerged for the research discussed to be taken further and in this section I mention some of the main opportunities and challenges that this work provides. I discuss future work in a number of sections, based on the contributions made by this thesis and the shortcomings that have been identified.

11.8.1 Affective Data with Discriminative Characteristics

In Chapter 6 I explored the application of affective feedback as an implicit source of evidence for topical relevance. A novel experimental framework was presented that allowed for the quantitative analysis and exploration of users' affective responses. The affective data that were gathered were used to train models capable of determining the topical relevance of viewed items. Overall, the performance of the models was found to be reasonable, though rather noisy. One potential explanation is that there is unnecessary redundancy, as well as overlap between different classes of values, which affects the modelling process. The experimental results described in Chapters 4, 6 and 7 suggest that a naturalistic setting that applies hidden recording would be in favour of acquiring spontaneous and authentic affective responses, thus producing data with more discriminative characteristics. Since the controlled experimental approach was not as effective as the naturalistic, it is important to isolate and mitigate those artificial factors that introduce noise.

There is also future work in developing an experimental framework that will promote engagement with the search task. A first step was to employ the simulated information need situations. Using the above framework I was able to represent the original content of the test collections in a way that facilitated a broader understanding of the search objectives and the information problem at hand. In addition, a layer of realism was added to the search tasks, while preserving well-defined relevance criteria. However, there is a need to redefine the context of the tasks with respect to a new set of goals, and enhance user motivation to achieve the highest correlation of emotions to task. This will result in a wider spectrum of affective behaviours and it will allow us to develop a more diverse collection of affective data. An avenue for future work would be to test the effect of a reward-punishment approach to the search tasks on user motivation. By rewarding or punishing users, based on their performance in locating relevant items, we can simulate real-life conditions and promote better engagement with the search task. Such framework would induce stronger emotional reactions, due to failure or success, that would result less noisy affective data and reduced overlap.

Another potential avenue for future work is the study of the relationship between the viewed content and users' affective responses towards it. The noisy performance of the affective models indicates that the content plays an important role, irrespective of its topical relevance, and can act as an emotional stimuli itself when linked, for example, to past experiences or memories. There is a need to explore ways of mitigating the influence that the content can exercise and reduce the amount of noise it introduces to the affective data.

11.8.2 Optimisation & Scope of Affective Models

The performance of the affective models presented in Chapter 6 suggests that affective feedback, as an implicit source of evidence for topical relevance, is a promising area of research. However, there were numerous modelling challenges and shortcomings identified throughout this thesis. First and foremost there is a need to examine additional classification techniques (linear discriminant analysis, HMM, neural networks, dynamic time warping, etc.) and seek alternative ways of viewing the data. The field of Machine Learning (ML) offers fundamental statistical-computational theories of learning processes, and has developed methods that are routinely used in many commercial applications and research areas. These can be applied to solve statistically complex problems or perform data-mining and uncover hidden attributes. Moreover, I examined a small set of modalities, which include facial expression analysis and physiological signal processing. The affective models could benefit from evaluating additional sensory channels (speech analysis, gesture recognition, electroencephalography, etc.) that could act in a complementary manner, per situation and per individual.

To emulate more closely user affective behaviour the models need to be trained to make decisions that resemble those of users, for different levels of topical relevance. In future work I intend to address this issue by training multi-layer hierarchical models, which as shown in Chapter 6 exhibit better performance. The training of weak classifiers and the application of information fusion, using weighted voting, is a more intelligent approach that will allow for superior classification rates. The weak classifiers will be trained to discriminate between affective responses for highly relevant, relevant, somewhat irrelevant and irrelevant information items; thus, modelling a range of affective responses that are associated with different degrees of topical relevance. Additional work is necessary in optimising and assessing the personalisation techniques over a greater number of individuals and a wider range of modalities. Finally, part of the accumulated evidence (behavioural, physiological, etc.) could be used to train models for assessing other aspects of the information seeking process, such as the difficulty of a search task, the levels of user engagement, the experienced anxiety, etc.

11.8.3 Interface Support

In Chapter 5 I tested an experimental video search interface that provided users with useful recommendations, after determining the relevance of the perused videos with the help of real-time facial expression analysis. The interface applied a layered-architecture, where each layer was dedicated to support a different facet of the search process. One of the main components of the proposed system was the affective model that processed data from a single modality and made relevance judgements. There is much scope for future work in developing an operational system that applies a bi- or multi-modal approach, allowing the employed modalities to complement each other and enhance the robustness of the system. More work is also necessary in ensuring the seemless integration of sensors and components, in a way that won't be perceived as intrusive.

In addition, the ethical and social dimensions of emotion recognition must be considered when employing affect-aware technologies in IR research. These critical concerns need to be carefully examined to avoid endangering the commercial and social acceptability of future affect-based systems. Finally, further work is necessary in dealing with scalability related issues. As pointed out by Joho, Jose, Valenti and Sebe (2009), the user-based approaches do not scale as much as content-based approaches do. Therefore, there is a need to explore ways of compensating for user-based information with content-based information (usually in the form of low level features), which can be more easily and unobtrusively extracted from the content.

11.9 Summary

This chapter has detailed opportunities to further the research presented in this thesis. The emerging affective paradigm presented here has fundamental implications for the design and evaluation of interactive IR systems. This work provides evidence of the importance of emotions in the information seeking process and the proposed affective feedback framework offers a means for using affective information to aid the retrieval process and improving the experience of a person searching. However, the affective paradigm needs to become broader and richer, and to encourage research that reaches across the boundaries of narrowly defined fields. In order to develop a holistic understanding of individual and context-based emotion behaviour future studies should employ a range of pertinent modalities that can imitate human sensory system, combined with content-based analysis of low-level features and high-level reasoning. The affective dimensions of information seeking behaviour, which have been overlooked in the past, can offer invaluable insights into making human-computer interactions more 'human". It is, therefore, vital that more work is undertaken to further this imaginative research and test these concepts in operational environments and longitudinal user experiments.

Bibliography

Aalbersberg, I. J.: 1992, Incremental relevance feedback, *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, pp. 11–22.

Adomavicius, G. and Member-Tuzhilin, A.: 2005, Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, *IEEE Transactions on Knowledge and Data Engineering* **17**(6), 734–749. Member-Adomavicius, Gediminas and Member-Tuzhilin, Alexander.

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

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

Ambady, N. and Rosenthal, R.: 1992, Thin slices of expressive behavior as predictors of interpersonal consequences : a meta-analysis, *Psychological Bulletin* **111**(2), 256–274.

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

Arapakis, I., Jose, J. M. and Gray, P. D.: 2008, Affective feedback: an investigation into the role of emotions in the information seeking process, *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, pp. 395–402.

Arapakis, I., Moshfeghi, Y., Joho, H., Ren, R., Hannah, D. and Jose, J. M.: 2009, Enriching user profiling with affective features for the improvement of a multimodal recommender system, *Proceeding of the ACM International Conference on Image and Video Retrieval*, ACM, New York, NY, USA, pp. 1–8.

Aue, T., Flykt, A. and Scherer, K. R.: 2007, First evidence of differential and sequential efferent effects of stimulus relevance and goal conduciveness appraisal, *Biological Psychology* **74**, 347–357.

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

Baeza-Yates, R. and Navarro, G.: 1996, Integrating contents and structure in text retrieval, *SIGMOD Rec.* **25**(1), 67–79.

Baeza-yates, R. and Ribeiro-Neto, B.: 1999, Modern information retrieval.

Bailey, P., Craswell, N. and Hawking, D.: 2003, Engineering a multi-purpose test collection for web retrieval experiments, *Information Processing and Management: an International Journal* **39**(6), 853–871.

Balomenos, T., Raouzaiou, A., Ioannou, S., Drosopoulos, A., Karpouzis, K. and Kollias, S.: 2005, *Machine Learning for Multimodal Interaction*, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, chapter Emotion Analysis in Man-Machine Interaction Systems, pp. 318–328.

Barrett, L. F. and Russell, J. A.: 1999, The structure of current affect: Controversies and emerging consensus, *Current Directions in Psychological Science* **8**(1), 10–14.

Bates, M. J.: 1989, The design of browsing and berrypicking techniques for the online search interface, *Online Information Review* **13**(5), 407–424.

Bates, M. J.: 1990, Where should the person stop and the information search interface start?, *Information Processing and Management: an International Journal* **26**(5), 575–591.

Beaulieu, M.: 1997, Experiments with interfaces to support query expansion, *Journal of Documentation* **53**(1), 8–19.

Beaulieu, M. and Jones, S.: 1998, Interactive searching and interface issues in the okapi best match probabilistic retrieval system, *Interacting with Computers* **10**(3), 237–248. HCI and Information Retrieval.

Belew, R. K.: 2001, *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW*, Cambridge University Press, New York, NY, USA.

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

Belkin, N. J., Cool, C., Head, J., Jeng, J., Kelly, D., Lin, S., Lobash, L., Park, S. Y., Savage-knepshield, P. and Sikora, C.: 2000, Relevance feedback versus local context analysis as term suggestion devices: Rutgers' trec-8 interactive track experience, *TREC-8, Proceedings of the Eighth Text Retrieval Conference*, Harman, pp. 565–574.

Belkin, N. J., Cool, C., Kelly, D., Lin, S.-J., Park, S. Y., Perez-Carballo, J. and Sikora, C.: 2001, Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval, *Information Processing and Management: an International Journal* **37**(3), 403–434.

Belkin, N. J., Cool, C., Stein, A. and Thiel, U.: 1995, Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems, *Expert Systems with Applications* **9**(3), 379–395.

Belkin, N. J. and Croft, W. B.: 1992, Information filtering and information retrieval: two sides of the same coin?, *Communications of the ACM* **35**(12), 29–38.

Bennett, J. L.: 1972, The user interface in interactive systems, *Annual Review of Information Science and Technology* **7**(159-196).

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

Billsus, D., Brunk, C. A., Evans, C., Gladish, B. and Pazzani, M.: 2002, Adaptive interfaces for ubiquitous web access, *Communications of the ACM* **45**(5), 34–38.

Billsus, D. and Pazzani, M. J.: 1999, A personal news agent that talks, learns and explains, *Proceedings of the third annual conference on Autonomous Agents*, ACM, New York, NY, USA, pp. 268–275.

Boone, R. T. and Cunningham, J. G.: 1998, Children's decoding of emotion in expressive body movement: The development of cue attunement, *Developmental Psychology* **34**(5), 1007–1016.

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

Boucsein, W.: 1992, *Electrodermal activity*, Plenum Press, New York.

Bower, G. H.: 1992, *The Handbook of Cognition and Emotion*, Hillsdale, New Jersey: Lawrence Erlbaum Associates, chapter How might emotions affect learning?, pp. 3–31.

Brave, S., Nass, C. and Hutchinson, K.: 2005, Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent, *International Journal of Human-Computer Studies* **62**(2), 161–178.

Breazeal, C.: 2001, *Designing Social Robots*, MIT Press, Cambridge, MA.

Brewster, S., Lumsden, J., Bell, M., Hall, M. and Tasker, S.: 2003, Multimodal 'eyes-free' interaction techniques for wearable devices, *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, New York, NY, USA, pp. 473–480.

Bruce, H. W.: 1994, A cognitive view of the situational dynamism of user-centered relevance estimation, *Journal of the American Society for Information Science* **45**(3), 142–148.

Buckley, C., Salton, G. and Allan, J.: 1994, The effect of adding relevance information in a relevance feedback environment, *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, Springer-Verlag New York, Inc., New York, NY, USA, pp. 292–300.

Bullington, J.: 2005, 'affective' computing and emotion recognition systems: the future of biometric surveillance?, *Proceedings of the 2nd annual conference on Information security curriculum development*, ACM, New York, NY, USA, pp. 95–99.

Campbell, I.: 2000, *The Ostensive Model of Developing Information-Needs*, PhD thesis, University of Glasgow.

Camurri, A., Lagerlöf, I. and Volpe, G.: 2003, Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques, *International Journal of Human-Computer Studies* **59**(1-2), 213–225.

Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S. and Karpouzis, K.: 2007, Multimodal emotion recognition from expressive faces, body gestures and speech, *Artificial Intelligence and Innovations 2007: from Theory to Applications*, Vol. 247, Springer-Verlag, Berlin, Heidelberg, pp. 375–388.

Castellano, G., Villalba, S. D. and Camurri, A.: 2007, Recognising human emotions from body movement and gesture dynamics, *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*, Springer-Verlag, Berlin, Heidelberg, pp. 71–82.

Cetintemel, U., Franklin, M. J. and Giles, C. L.: 2000, Self-adaptive user profiles for large-scale data delivery, *Proceedings of the 16th International Conference on Data Engineering*, IEEE Computer Society, Washington, DC, USA, p. 622.

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

Chandra, A. and Calderon, T.: 2005, Challenges and constraints to the diffusion of biometrics in information systems, *Communications of the ACM* **48**(12), 101–106.

Chanel, G., Rebetez, C., Bétrancourt, M. and Pun, T.: 2008, Boredom, engagement and anxiety as indicators for adaptation to difficulty in games, *Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era*, ACM, New York, NY, USA, pp. 13–17.

Chen, F.-S., Fu, C.-M. and Huang, C.-L.: 2003, Hand gesture recognition using a real-time tracking method and hidden markov models, *Image and Vision Computing* **21**(8), 745–758.

Christiansen, A. and Lee, D. J.: 2006, Relevance feedback query refinement for pdf medical journal articles, *Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems*, IEEE Computer Society, Washington, DC, USA, pp. 57–62.

Claypool, M., Le, P., Wased, M. and Brown, D.: 2001, Implicit interest indicators, *Proceedings of the 6th international conference on Intelligent user interfaces*, ACM, New York, NY, USA, pp. 33–40.

Cohn, J. and Kanade, T.: 2006, Use of automated facial image analysis for measurement of emotion expression, *in* J. A. Coan and J. B. Allen (eds), *The Handbook of Emotion Elicitation and Assessment*, Oxford University Press Series in Affective Science.

Cornelius, R. R.: 2000, Theoretical approaches to emotion, *Proceedings of ISCA: workshop on Speech and Emotion* pp. 3–11.

Corradini, A., Mehta, M., Bernsen, N. and Martin, J. C.: 2003, Multimodal input fusion in human–computer interaction, *NATO-ASI Conference on Data Fusion for Situation Monitoring, Incident Detection, Alert, and Response Management*.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. and Taylor, J. G.: 2001, Emotion recognition in human-computer interaction, *Signal Processing Magazine, IEEE* **18**(1), 32–80.

Croft, W. B.: 1993, Knowledge-based and statistical approaches to text retrieval, *IEEE Expert: Intelligent Systems and Their Applications* **8**(2), 8–12.

Croft, W. B. and Thompson, R. H.: 1987, I3r: a new approach to the design of document retrieval systems, *Journal of the American Society for Information Science* **38**(6), 389–404.

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

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

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

Darwin, C.: 2005, *The Expression of the Emotions in Man and Animals*, Kessinger Publishing.

Denoyer, L. and Gallinari, P.: 2006, The Wikipedia XML Corpus, *SIGIR Forum* .

Dickie, C., Hart, J., Vertegaal, R. and Eiser, A.: 2006, Lookpoint: an evaluation of eye input for hands-free switching of input devices between multiple computers, *Proceedings of the 18th Australia conference on Computer-Human Interaction*, ACM, New York, NY, USA, pp. 119–126.

Dickie, C., Vertegaal, R., Fono, D., Sohn, C., Chen, D., Cheng, D., Shell, J. S. and Aoudeh, O.: 2004, Augmenting and sharing memory with eyeblog, *Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, ACM, New York, NY, USA, pp. 105–109.

Duchowski, A. T.: 2002, A breadth-first survey of eye-tracking applications., *Behavior Research Methods, Instruments & Computers* **34**(4), 455–470.

Ekman, P.: 1984, *Approaches to Emotion*, Hillsdale, New Jersey: Lawrence Erlbaum Associates, chapter Expression and the nature of emotion, pp. 319–344.

Ekman, P.: 1992, An argument for basic emotions, *Cognition and Emotion* **6**, 169–200.

Ekman, P.: 1999a, *Basic Emotions*, The Handbook of Cognition and Emotion, U.K.: John Wiley & Sons, Ltd, chapter 6, pp. 45–60.

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

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

Ekman, P.: 2003b, *Unmasking the face*, MA: Malor books, Cambridge.

Ekman, P. and Friesen, W. V.: 1975, *Unmasking the face. A guide to recognizing emotions from facial clues*, Englewood Cliffs, New Jersey: Prentice-Hall.

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

Essa, I. A.: 1995, *Analysis, interpretation and synthesis of facial expressions*, PhD thesis, Cambridge, MA, USA.

Essa, I. A. and Pentland, A. P.: 1997, Coding, analysis, interpretation, and recognition of facial expressions, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**(7), 757–763.

Fasel, B. and Luettin, J.: 2003, Automatic facial expression analysis: A survey, *Pattern Recognition* **36**(1), 259–275.

Folkman, S., Lazarus, R. S., Gruen, R. J. and DeLongis, A.: 1986, Appraisal, coping, health status, and psychological symptoms, *Journal of Personality and Social Psychology* **50**(3), 571–579.

Frijda, N. H.: 1993, Appraisal and beyond: The issue of cognitive determinants of emotion, *Cognition and Emotion* **7**(3 & 4), 357–387.

Frijda, N. H., Manstead, A. S. R. and Bem, S. (eds): 2000, *Emotions and Beliefs: How Feelings Influence Thoughts*, Cambridge University Press.

Furnas, G. W. and Russell, D. M.: 2005, Making sense of sensemaking, *CHI '05 extended abstracts on Human factors in computing systems*, ACM, New York, NY, USA, pp. 2115–2116.

Go, H.-J., Kwak, K.-C., Lee, D.-J. and Chun, M.-G.: 2003, Emotion recognition from the facial image and speech signal, Vol. 3, pp. 2890–2895.

Gunes, H. and Piccardi, M.: 2007, Bi-modal emotion recognition from expressive face and body gestures, *Journal of Network and Computer Applications* **30**(4), 1334–1345.

Gwizdka, J. and Lopatovska, I.: 2009, The role of subjective factors in the information search process. Manuscript submitted for publication.

Han, E.-H. and Karypis, G.: 2005, Feature-based recommendation system, *Proceedings of the 14th ACM international conference on Information and knowledge management*, ACM, New York, NY, USA, pp. 446–452.

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

Hansen, D. W., Hansen, J. P., Niels, M. and Stegmann, M. B.: 2002, Eye typing using markov and active appearance models, *Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision*, IEEE Computer Society, Washington, DC, USA, p. 132.

Harman, D.: 1992, Relevance feedback revisited, *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, pp. 1–10.

Healey, J., Seger, J. and Picard, R.: 1999, Quantifying driver stress: Developing a system for collecting and processing bio-metric signals in natural situations, *Processing Biometric Signals in Natural Situations", Proceedings of the Rocky Mountain Bio-Engineering Symposium, 16 th - 18 th*, M.I.T. Media Laboratory.

Heishman, R., Duric, Z. and Wechsler, H.: 2004, Using eye region biometrics to reveal affective and cognitive states, *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 5*, IEEE Computer Society, Washington, DC, USA, p. 69.

Hiemstra, D.: 2000, *Using Language Models for Information Retrieval*, PhD thesis, Centre for Telematics and Information Technology.

Hsu, C.-W., Chang, C.-C. and Lin, C.-J.: 2003, A practical guide to support vector classification, *Technical report*, Department of Computer Science and Information Engineering, National Taiwan University.

Ingwersen, P.: 1992, *Information Retrieval Interaction*, Taylor Graham.

Ingwersen, P.: 1994, Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction, *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, Springer-Verlag New York, Inc., New York, NY, USA, pp. 101–110.

Ingwersen, P.: 1996, Cognitive perspectives of information retrieval interaction: elements of a cognitive ir theory, *Journal of Documentation* **52**, 3–50.

Isbister, K., Höök, K., Sharp, M. and Laaksolahti, J.: 2006, The sensual evaluation instrument: developing an affective evaluation tool, *Proceedings of the SIGCHI conference on Human Factors in computing systems*, ACM, New York, NY, USA, pp. 1163–1172.

Jacob, R. J. K.: 1995, Eye tracking in advanced interface design, pp. 258–288.

Jaimes, A. and Sebe, N.: 2007, Multimodal human-computer interaction: A survey, *Computer Vision and Image Understanding* **108**(1-2), 116–134.

Jain, A. K. and Ross, A.: 2004, Multibiometric systems, *Communications of the ACM*, Vol. 47, pp. 34–40.

Jameson, A., Schäfer, R., Weis, T., Berthold, A. and Weyrath, T.: 1999, Making systems sensitive to the user's time and working memory constraints, *Proceedings of the 4th international conference on Intelligent user interfaces*, ACM, New York, NY, USA, pp. 79–86.

Järvelin, K.: 2007, An analysis of two approaches in information retrieval: From frameworks to study designs, *Journal of the American Society for Information Science and Technology* **58**(7), 971–986.

Joachims, T., Freitag, D. and Mitchell, T.: 1997, Webwatcher: A tour guide for the world wide web, *In Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, pp. 770–775.

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

Joho, H., Jose, J. M., Valenti, R. and Sebe, N.: 2009, Exploiting facial expressions for affective video summarisation, *Proceeding of the ACM International Conference on Image and Video Retrieval*, ACM, New York, NY, USA, pp. 1–8.

Julien, H., McKechnie, L. E. F. and Hart, S.: 2005, Affective issues in library and information science systems work: A content analysis, *Library & Information Science Research* **27**(4), 453–466.

Kahneman, D.: 2000, *Choices, Values, and Frames*, Cambridge University Press, New York, NY, USA, chapter Experienced utility and objective happiness: A moment-based approach, pp. 673–692.

Kaliouby, R. E. and Robinson, P.: 2004, Real-time inference of complex mental states from facial expressions and head gestures, *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 10*, IEEE Computer Society, Washington, DC, USA, p. 154.

Kapoor, A., Picard, R. W. and Ivanov, Y.: 2004, Probabilistic combination of multiple modalities to detect interest, *Proceedings of the 17th International Conference on Pattern Recognition, 2004*, Vol. 3, pp. 969–972.

Karat, J.: 2003, Beyond task completion: evaluation of affective components of use, pp. 1152–1164.

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

Kelly, D. and Belkin, N. J.: 2001, Reading time, scrolling and interaction: exploring implicit sources of user preferences for relevance feedback, *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, pp. 408–409.

Kelly, D. and Belkin, N. J.: 2002, A user modeling system for personalized interaction and tailored retrieval in interactive ir, *Proceedings of the American Society for Information Science and Technology* **39**(1), 316–325.

Kelly, D. and Cool, C.: 2002, The effects of topic familiarity on information search behavior, *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, ACM, New York, NY, USA, pp. 74–75.

Kelly, D. and Teevan, J.: 2003, Implicit feedback for inferring user preference: a bibliography, *SIGIR Forum* **37**(2), 18–28.

Kelly, J. D.: 2004, *Understanding implicit feedback and document preference: a naturalistic user study*, PhD thesis, New Brunswick, NJ, USA.

Kim, K.-S.: 2008, Effects of emotion control and task on web searching behavior, *Information Processing and Management: an International Journal* **44**(1), 373–385.

Kirishima, T., Sato, K. and Chihara, K.: 2005, Real-time gesture recognition by learning and selective control of visual interest points, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(3), 351–364.

Kleinginna, P. R. and Kleinginna, A. M.: 2005, A categorized list of motivation definitions, with a suggestion for a consensual definition, *Motivation and Emotion* **5**(3), 263–291.

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

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

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, *Communications of the ACM* **40**(3), 77–87.

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

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

Larsen, R. J. and Fredrickson, B. L.: 1999, *Well-being: The foundations of hedonic psychology*, New York: Russell Sage Foundation, chapter Measurement issues in emotion research, pp. 40–60.

Lavie, T. and Tractinsky, N.: 2004, Assessing dimensions of perceived visual aesthetics of web sites, *International Journal of Human-Computer Studies* **60**(3), 269–298.

Lazarus, R. S.: 1984, *Approaches to Emotion*, Hillsdale, New Jersey: Lawrence Erlbaum Associates, chapter Thoughts on the relations between emotion and cognition, pp. 247–259.

Lee, H. J. and Neal, D.: 2007, Towards web 2.0 music information retrieval: utilizing emotion-based, user-assigned descriptions, *Proceedings of the 70th Annual Meeting of the American Society of Information Science and Technology*, Vol. 45 of *Joining Research and Practice: social computing and information science*, pp. 732–741.

Licsár, A. and Szirányi, T.: 2005, User-adaptive hand gesture recognition system with interactive training, *Image and Vision Computing* **23**(12), 1102–1114.

Lin, H.-T. and Lin, C.-J.: 2003, A study on sigmoid kernels for svm and the training of non-psd kernels by smo-type methods, *Technical report*, National Taiwan University.

Linden, G., Smith, B. and York, J.: 2003, Amazon.com recommendations: item-to-item collaborative filtering, *Internet Computing, IEEE* **7**(1), 76–80.

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

Lopatovska, I.: 2009, Searching for good mood: examining relationships between search task and mood, *Proceedings of the 72th Annual Meeting of the American Society for Information Science and Technology*. Manuscript submitted for publication.

Lopatovska, I. and Cool, C.: 2008, Online search: Uncovering affective characteristics of information retrieval experience. Presented at the 2008 ALISE Annual Conference, Philadelphia, PA.

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

MacKay, D. M.: 1969, *Information mechanism and meaning*, The M.I.T. Press.

Marchionini, G.: 2004, From information retrieval to information interaction, *In Proceedings of the 26th European Conference on IR Research*, pp. 1–11.

McAllister, G., McKenna, S. J. and Ricketts, I. W.: 2002a, Hand tracking for behaviour understanding, *Image and Vision Computing* **20**(12), 827–840.

McAllister, G., McKenna, S., J. and Ricketts, I. W.: 2002b, Hand tracking for behaviour understanding, *Image and Vision Computing* **20**(12), 827–840.

McKechnie, L., S., R. C. and Rothbauer, P.: 2007, *Affective dimensions of information seeking in the context of reading*, Medford, NJ: Information Today.

Meijer, M.: 2005, The contribution of general features of body movement to the attribution of emotions, *Journal of Nonverbal Behavior* **13**(4), 247–268.

Mendoza, E. and Carballo, G.: 1999, Vocal tremor and psychological stress, *Journal of Voice* **13**(1), 105–112.

Miller, B. N., Albert, I., Lam, S. K., Konstan, J. A. and Riedl, J.: 2003, Movielens unplugged: experiences with an occasionally connected recommender system, *Proceedings of the 8th international conference on Intelligent user interfaces*, ACM, New York, NY, USA, pp. 263–266.

Mitroff, I. I.: 1974, A brunswik lens model of dialectical inquiring systems, *Journal of Theory and Decision* **5**(1), 45–67.

Mooers, C. N.: 1950, Information retrieval viewed as temporal signaling, *Proceedings of the International Congress of Mathematicians*, Vol. 1, pp. 572–573.

Mooney, C., Scully, M., Jones, G. J. F. and Smeaton, A. F.: 2006, Investigating biometric response for information retrieval applications, *Lecture Notes in Computer Science* pp. 570–574.

Moore, M. M. and Dua, U.: 2004, A galvanic skin response interface for people with severe motor disabilities, *ACM SIGACCESS Accessibility and Computing* (77-78), 48–54.

Morency, L.-P. and Darrell, T.: 2006, Head gesture recognition in intelligent interfaces: the role of context in improving recognition, *Proceedings of the 11th international conference on Intelligent user interfaces*, ACM, New York, NY, USA, pp. 32–38.

Morency, L.-P., Sidner, C., Lee, C. and Darrell, T.: 2007, Head gestures for perceptual interfaces: The role of context in improving recognition, *Artificial Intelligence* **171**(8-9), 568–585.

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

Murray, I. R. and Arnott, J. L.: 1993, Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion, *Journal of the Acoustic Society of America* **93**(2), 1097–1108.

Nahl, D.: 1998a, Ethnography of novices' first use of web search engines: affective control in cognitive processing, *Internet Reference Services Quarterly* **3**(2), 51–72.

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

Nahl, D.: 2004, Measuring the affective information environment of web searchers, *Proceedings of the American Society for Information Science and Technology*, Vol. 41, pp. 191–197.

Nahl, D.: 2005, Affective and cognitive information behavior: Interaction effects in internet use, *Proceedings of the American Society for Information Science and Technology*, Vol. 42.

Nahl, D. and Bilal, D. (eds): 2007, *Information and Emotion: the Emergent Affective Paradigm in Information Behavior Research and Theory*, Medford, New Jersey: Information Today, Inc.

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

Ng, C. W. and Ranganath, S.: 2002, Real-time gesture recognition system and application, *Image and Vision Computing* **20**(13-14), 993–1007.

Nichols, D. M.: 1997, Implicit rating and filtering, *In Proceedings of the Fifth DELOS Workshop on Filtering and Collaborative Filtering*, pp. 31–36.

Oard, D. W. and Kim, J.: 2001, Modeling information content using observable behavior.

Odlyzko, A. M.: 2003, Internet traffic growth: Sources and implications, *Proceedings of the 8th International Conference on Intelligent User Interfaces*, pp. 263–266.

Pantic, M. and Patras, I.: 2006, Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences, *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* **36**(2), 433–449.

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

Pantic, M. and Rothkrantz, L. J. M.: 2000b, Automatic analysis of facial expressions: the state of the art, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(12), 1424–1445.

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

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

Pelz, J. B., Canosa, R. L., Kucharczyk, D., Babcock, J. S., Silver, A. and Konno, D.: 2000, Portable eyetracking: A study of natural eye movements, *in* B. E. Rogowitz and T. N. Pappas (eds), *Human Vision and Electronic Imaging V*, SPIE Proceedings, p. 3659.

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.

Picard, R. W.: 1997, *Affective computing*, MIT Press, Cambridge, MA, USA.

Picard, R. W.: 2001, Building hal: Computers that sense, recognize, and respond to human emotion, *Society of Photo-Optical Instrumentation Engineers*, Vol. 4299 of *Human Vision and Electronic Imaging VI*, pp. 518–523.

Picard, R. W.: 2003, Affective computing: challenges, *International Journal of Human-Computer Studies* **59**(1-2), 55–64.

Picard, R. W. and Daily, S. B.: 2005, Evaluating affective interactions: Alternatives to asking what users feel, *CHI 2005 Workshop 'Evaluating Affective Interfaces'*.

Picard, R., Wexelblat, A. and Nass, C. I.: 2002, Future interfaces: social and emotional, *CHI '02 extended abstracts on Human factors in computing systems*, pp. 698–699.

Plutchik, R.: 1980, *Emotion: Theory, research and experience*, Vol. 1 of *Theories of emotion*, New York: Academic, chapter A general psychoevolutionary theory of emotion, pp. 3–33.

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

Reeves, B. and Nass, C. I.: 1996, *The Media Equation: How People Treat Computers, Television and New Media Like Real People and Places (Book Review)*, Cambridge University Press, Cambridge, England.

Reynolds, C. and Picard, R. W.: 2004, Affective sensors, privacy, and ethical contracts, *CHI '04 extended abstracts on Human factors in computing systems*, ACM, New York, NY, USA, pp. 1103–1106.

Robertson, S. E. and Hancock-Beaulieu, M. M.: 1992, On the evaluation of ir systems, *Information Processing and Management: an International Journal* **28**(4), 457–466.

Rosenberg, E. L. and Ekman, P.: 1993, *Facial Expression and Emotion*, Neuroscience Year: Supplement 3 to the Encyclopedia of Neuroscience, Birkhäuser, pp. 51–52.

Rothkopf, C. A. and Pelz, J. B.: 2004, Head movement estimation for wearable eye tracker, *Proceedings of the 2004 symposium on Eye tracking research & applications*, ACM, New York, NY, USA, pp. 123–130.

Rui, Y. and Huang, T.: 2000, Optimizing learning in image retrieval, *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 236–243.

Russell, J. A.: 1994, Is there universal recognition of emotion from facial expression?, *Psychological Bulletin* **115**, 102–141.

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

Russell, J. A. and Mehrabian, A.: 1977, Evidence for a three-factor theory of emotions, *Journal of Research in Personality* **11**(3), 273–294.

Russell, J. A. and Steiger, J. H.: 1982, The structure in persons' implicit taxonomy of emotions, *Research in Personality* **16**, 447–469.

Salojärvi, J., Kojo, I., Jaana, S. and Kaski, S.: 2003, Can relevance be inferred from eye movements in information retrieval?, *Proceedings of the Workshop on Self-Organizing Maps*, pp. 261–266.

Salojärvi, J., Puolamäki, K., Simola, J., Kovanen, L., Kojo, I. and Kaski, S.: 2005, Inferring relevance from eye movements: Feature extraction, *Publications in computer and information science, report a82*, Helsinki University of Technology.

Salton, G.: 1968, A comparison between manual and automatic indexing methods, *Technical report*, Ithaca, NY, USA.

Salton, G. and Buckley, C.: 1990, Improving retrieval performance by relevance feedback, *Journal of the American Society for Information Science* **41**(4), 288–297.

Salton, G. and Buckley, C.: 1997, Improving retrieval performance by relevance feedback, pp. 355–364.

Sander, D., Grandjean, D., Pourtois, G., Schwartz, S., Seghier, M. L., Scherer, K. R. and Vuilleumier, R.: 2005, Emotion and attention interactions in social cognition: Brain regions involved in processing anger prosody, *NeuroImage* **28**(4), 848–858. Special Section: Social Cognitive Neuroscience.

Sander, D., Grandjean, D. and Scherer, K. R.: 2005, A systems approach to appraisal mechanisms in emotion, *Neural Networks* **18**(4), 317–352.

Saracevic, T.: 1975, Relevance: A review of and a framework for the thinking on the notion in information science, *Journal of the American Society for Information Science* **26**(6), 321–343.

Saracevic, T.: 1996, Relevance reconsidered, *Proceedings of the Second Conference on Conception of Library and Information Science*, pp. 201–218.

Saracevic, T.: 2007, Relevance: A review of the literature and a framework for thinking on the notion in information science. part iii: Behavior and effects of relevance, *Journal of the American Society for Information Science and Technology* **58**(13), 2126–2144.

Savino, P. and Sebastiani, F.: 1998, Essential bibliography on multimedia information retrieval, categorisation and filtering., *Slides of the 2nd European Digital Libraries Conference Tutorial on Multimedia Information Retrieval.*

Schapira, E. and Sharma, R.: 2001, Experimental evaluation of vision and speech based multimodal interfaces, *Proceedings of the 2001 workshop on Perceptive user interfaces*, ACM, New York, NY, USA, pp. 1–9.

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

Scherer, K. R.: 2002, *Emotion, the psychological structure of emotions*, International Encyclopedia of the Social & Behavioral Sciences, Harvard Libraries, Oxford.

Scherer, K. R.: 2003, Vocal communication of emotion: a review of research paradigms, *Speech Commun.* **40**(1-2), 227–256.

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

Schmidt, S. and Stock, W. G.: 2009, Collective indexing of emotions in images. a study in emotional information retrieval, *Journal of the American Society for Information Science and Technology* **60**(5), 863–876.

Schuller, B., Arsic, D., Wallhoff, F. and Rigoll, G.: 2006, Emotion recognition in the noise applying large acoustic feature sets., *Proceedings of Speech Prosody*, International Symposium on Computer Architecture, 2006, Dresden.

Schuller, B., Lang, M. and Rigoll, G.: 2002, Multimodal emotion recognition in audiovisual communication, Vol. 1, pp. 745–748.

Schutz, A.: 1970, *Reflections on the problem of relevance*, Yale University Press.

Schwarzkopf, E.: 2004, Enhancing the interaction with information portals, *Proceedings of the 9th international conference on Intelligent user interfaces*, ACM, New York, NY, USA, pp. 322–324.

Sebe, N., Bakker, E., Cohen, I., Gevers, T. and Huang, T. S.: 2005, Bimodal emotion recognition, *5th International Conference on Methods and Techniques in Behavioral Research*, Wageningen, Netherlands.

Sebe, N., Cohen, I. and Huang, T. S.: 2005, *Multimodal Emotion Recognition*, Handbook of Pattern Recognition and Computer Vision, World Scientific.

Sebe, N., Lew, M., S., Cohen, I., Sun, Y., Gevers, T. and Huang, T. S.: 2004, Authentic facial expression analysis, pp. 517–522.

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.

Sensuse, D. I.: 2004, *A comparison of manual indexing and automatic indexing in the Humanities*, PhD thesis, University of Toronto.

Seo, Y.-W. and Zhang, B.-T.: 2000, Learning user's preferences by analyzing web-browsing behaviors, *Proceedings of the fourth international conference on Autonomous agents*, ACM, New York, NY, USA, pp. 381–387.

Shannon, C. E.: 2001, A mathematical theory of communication, *ACM SIGMOBILE Mobile Computing and Communications Review* **5**(1), 3–55.

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

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

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

Smith, J. D., Vertegaal, R. and Sohn, C.: 2005, Viewpointer: lightweight calibration-free eye tracking for ubiquitous handsfree deixis, *Proceedings of the 18th annual ACM symposium on User interface software and technology*, ACM, New York, NY, USA, pp. 53–61.

Snelick, R., Indovina, M., Yen, J. and Mink, A.: 2003, Multimodal biometrics: issues in design and testing, *Proceedings of the 5th international conference on Multimodal interfaces*, ACM, New York, NY, USA, pp. 68–72.

Soleymani, M., Chanel, G., Kierkels, J. J. M. and Pun, T.: 2008, Affective ranking of movie scenes using physiological signals and content analysis, *Proceeding of the 2nd ACM workshop on Multimedia semantics*, ACM, New York, NY, USA, pp. 32–39.

Song, M., Bu, J., Chen, C. and Li, N.: 2004, Audio-visual based emotion recognition - a new approach, Vol. 2, pp. 1020–1025.

Spink, A.: 1994, Term relevance feedback and query expansion: relation to design, *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, Springer-Verlag New York, Inc., New York, NY, USA, pp. 81–90.

Spink, A., Greisdorf, H. and Bateman, J.: 1998, From highly relevant to not relevant: examining different regions of relevance, *Information Processing and Management: an International Journal* **34**(5), 599–621.

Swindells, C., MacLean, K. E., Booth, K. S. and Meitner, M.: 2006, A case-study of affect measurement tools for physical user interface design, *Proceedings of Graphics Interface 2006*, Canadian Information Processing Society, Toronto, Ont., Canada, Canada, pp. 243–250.

Tamine, L., Boughanem, M. and Zemirli, N.: 2006, Inferring the user interests using the search history, *in* K. D. Althoff and M. Schaaf (eds), *LWA*, Vol. 1/2006 of *Hildesheimer Informatik-Berichte*, University of Hildesheim, Institute of Computer Science, pp. 108–110.

Tan, K.-H., Kriegman, D. J. and Ahuja, N.: 2002, Appearance-based eye gaze estimation, *Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision*, IEEE Computer Society, Washington, DC, USA, p. 191.

Tang, R. and Solomon, P.: 1998, Toward an understanding of the dynamics of relevance judgement: an analysis of one person's search behavior, *Information Processing and Management: an International Journal* **34**(2-3), 237–256.

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

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

Tomkins, S. S.: 1984, *Approaches to Emotion*, Hillsdale, New Jersey: Lawrence Erlbaum Associates, chapter Affect theory, pp. 163–197.

Tomkins, S. S.: 1995, *Exploring Affect: The Selected Writtings of Sylvan S. Tomkins*, New York: Cambridge University Press.

Tsalakanidou, F., Malassiotis, S. and Strintzis, M., G.: 2007, A 3d face and hand biometric system for robust user-friendly authentication, *Pattern Recognition Letters* **28**(16), 2238–2249.

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

Vertegaal, R., Dickie, C., Sohn, C. and Flickner, M.: 2002, Designing attentive cell phone using wearable eyecontact sensors, *CHI '02 extended abstracts on Human factors in computing systems*, ACM, New York, NY, USA, pp. 646–647.

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

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

Wallbott, H. G.: 1998, Bodily expression of emotion, *European Journal of Social Psychology* **28**(6).

Wang, J.-G., Sung, E. and Venkateswarlu, R.: 2003, Eye gaze estimation from a single image of one eye, *Proceedings of the Ninth IEEE International Conference on Computer Vision*, IEEE Computer Society, Washington, DC, USA, p. 136.

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

Wang, Y., Ai, H., Wu, B. and Huang, C.: 2004, Real time facial expression recognition with adaboost, Vol. 3, pp. 926–929.

Ware, C. and Mikaelian, H. H.: 1987, An evaluation of an eye tracker as a device for computer input2, *ACM SIGCHI Bulletin* **18**(4), 183–188.

Watson, D., Clark, L. A. and Tellegen, A.: 1988, Development and validation of brief measures of positive and negative affect: The panas scales, *Journal of Personality and Social Psychology* **54**(6), 1063–1070.

Weiser, M.: 1993, Hot topics-ubiquitous computing, *Computer* **26**(10), 71–72.

Weiser, M.: 1994, Ubiquitous computing, *ACM Conference on Computer Science*, p. 418.

White, R. W.: 2004, *Implicit feedback for interactive information retrieval*, PhD thesis, University of Glasgow.

White, R. W., Ruthven, I. and Jose, J. M.: 2002, Finding relevant documents using top ranking sentences an evaluation of two alternative schemes, *In Proceedings of SIGIR 2002*, ACM Press, pp. 57–64.

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

Yoshitomi, Y., Sung-Ill, K., Kawano, T. and Kilazoe, T.: 2000, Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face, pp. 178–183.

Zajonc, R. B.: 1984, *Approaches to Emotion*, Hillsdale, New Jersey: Lawrence Erlbaum Associates, chapter The interaction of affect and cognition, pp. 239–246.

Zeng, Z., Hu, Y., Fu, Y., Huang, T. S., Roisman, G. I. and Wen, Z.: 2006, Audio-visual emotion recognition in adult attachment interview, *Proceedings of the 8th international conference on Multimodal interfaces*, ACM, New York, NY, USA, pp. 139–145.

Zhou, X. S. and Huang, T. S.: 2001, Exploring the nature and variants of relevance feedback, *Proceedings of the IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'01)*, IEEE Computer Society, Washington, DC, USA, p. 94.

Zhu, Z., Fujimura, K. and Ji, Q.: 2002, Real-time eye detection and tracking under various light conditions, *Proceedings of the 2002 symposium on Eye tracking research & applications*, ACM, New York, NY, USA, pp. 139–144.